CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Analysis of expression quantitative trait loci in D. melanogaster
Authors	Harnett, Dermot
Publication date	2016
Original Citation	Harnett, D. 2016. Analysis of expression quantitative trait loci in D. melanogaster. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2016, Dermot Harnett. - http://creativecommons.org/licenses/by-nc-nd/3.0/
Download date	2023-05-05 20:50:52
Item downloaded from	http://hdl.handle.net/10468/3114

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Analysis of Expression Quantitative Trait Loci in *D. melanogaster*

Volume 1 of 1

Dermot Harnett B.A. Sc Human Genetics

Submitted to University College Cork

Research Conducted in: The Department of Genome Biology,

European Molecular Biology Laboratory, Heidelberg

Submission Date: 13 May 2016

Department Head: Eileen Furlong

Research Supervisor: Eileen Furlong

TABLE OF CONTENTS

I, Dermot Harnett, certify that this thesis is my own work and I have not obtained a degree in this university or elsewhere on the basis of the work submitted in this thesis.

Dermot Harnett

LIST OF FIGURES

List of Abbreviations and Acronyms

3iQTL	3’ Isoform Quantitative Trait Locus
CAGE	Cap Analysis Gene Expression
CRM	*Cis* Regulatory Module
DBP	DNA-binding Protein
DGRP	*Drosophila* Genetic Reference Panel
DHS	DNase Hypersensitivity Site
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
GRO-seq	Global Run On Sequencing
pA site	polyadenylation site
PRO-seq	Precision nuclear Run-On and sequencing assay
PBM	Protein Binding Microarray
PWM	Position Weight Matrix
RBP	RNA Binding Protein
RNA	RiboNucleic Acid
RNAP	RNA Polymerase
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TPM	Tags Per Million
tssQTL	Transcription Start Site Quantitative Trait Locus
TSS	Transcription Start Site

A note on publications:

The work on 3'UTR biology, polyadenylation site motifs, and 3'-Tag-seq eQTL in this Thesis has been resubmitted following review to the Journal Nature, as part of a paper entitled "Genetic and developmental regulation of expression levels and isoform diversity during embryogenesis" in which I am the second author.

The work on CAGE peak calling, TSS associated motifs, and tssQTL in this paper forms part of a manuscript entitled "Genetics variation uncovers a link between promoter shape, genetic robustness and expression noise" and has been submitted for publication. I am a third author on this paper.

The work on enhancer transcription will form part of a paper on enhancer transcription in *Drosophila* , in which I will be first author.

Acknowledgements:
I would like to acknowledge my many collaborators, who made this work possible:

Enrico Cavanno – for carrying out embryo collections, and most experimental work in the 3' Tag Seq sections of this thesis, and for editing of the figures in chapter 4.

Nils Koelling – for carrying out the read processing, polyadenylation site calls, and eQTL calling on the 3' Tag seq sections of this thesis editing of figures in chapter 4.

Ignacio Schor– for performing the experimental work in the 5' CAGE sections of this thesis, and editing of figures in chapter 4.

Olga Mikhalylichenko – for performing the experimental work in the enhancer transcription related sections of this thesis, editing of figures in chapter 2.

Vladislav Bondarenk- for various computational tasks in the enhancer transcription related sections of this thesis, including quality checking of datasets and enhancer sets.

I would like to thank my supervisor, Eileen Furlong, for making this work possible, for her input and support, and for saying what I needed to hear sometimes even when I didn't want to hear it. I would also like to thank David Garfield for providing so much intellectual and emotional support these past four years.

Dad, I don't know who I'd be or what I'd have done without you. Everything good I've managed has been with you behind me.

Mary, Niamh, Brian, Kieran, Helen, I've missed you terribly.

Maja, you were a saint. Thank you for your patience and kindness.

And to my friends in Ireland and abroad, thank you, and all glory to the SlackMind.

Abstract

Expression Quantitative Trait Loci (eQTL) analysis allows for the identification of genetic variation associated with variation in gene expression. It is often unclear however, which of the associated variants are causal, and by what mechanism. Integrating functional genomic data with eQTL data can provide insight into the impact of natural variation in the population, and the nature of the transcriptional machinery itself. In this thesis, I integrate functional genomic data with eQTL data derived from both 5' CAGE and 3' Tag-seq expression assays, in developing embryos. I first use both datasets to analyze the transcription landscape in embryonic *D. melanogaster*, and then carry out an analysis of sequence motifs associated with transcription factor binding sites, promoters, and 3' polyadenylation sites. Finally, I integrate functional genomic data, including these novel sequence motifs, to shed light on the mechanisms of gene expression variation in *D. melanogaster*. I am able to demonstrate that some variants effecting gene regulation in *Drosophila* are found within haplotypes which buffer their effects.

1 Introduction

1.1 Overview: Mapping genotype to phenotype

The link between genotype and phenotype is now relatively clear for protein coding sequence – the Genetic Code and the Central Dogma of how it is interpreted, are the foundations of modern molecular biology. But genomes contain more than protein coding sequence. The development of a fertilized zygote into a complex, multicellular organism requires that the correct cells produce RNA and proteins at the correct time, in a tightly controlled manner. The genome thus contains information not just on 'what' - but on 'where' 'when' and 'how much' – the fundamental properties or logic inherent in gene regulation. This body of information differs markedly from coding sequence: despite decades of searching for a 'regulatory code', it has become clear that there is no neat mapping between DNA sequence and spatiotemporal phenotypes as there is from to DNA to RNA or protein sequence.

No single protein complex or system, like the ribosome, translates spatiotemporal information. Rather, gene expression is the result of the many processes that lead from transcription of a viable mRNA to its translation at the ribosome, and turnover of the resulting protein. Transcription initiation itself is controlled directly and indirectly by the presence of sequence motifs that bind the core transcriptional machinery, or other regulatory proteins such as transcription factors. CRMs (Cis Regulatory Modules) - sequences that bind binding regulatory proteins – integrate the input of different factors, and allow specific sequences to be transcribed in response to specific cellular conditions. Transcripts are transcribed and processed from promoters, released from DNA, transported from the nucleus, translated and degraded, and each of these steps are also regulated by specific DNA and RNA elements, such as promoter motifs, and 3' cleavage motifs.

Furthermore, unlike the inherently sequential and modular nature of protein assembly, spatiotemporal control of gene expression often depends on the simultaneous, analog interaction of numerous components. The 'regulatory code'

then, if it can be said to exist, is a complicated one, and differs significantly between organisms and cell types. Nevertheless, our understanding of this information is of crucial importance. In it is in large part due to differences in gene expression that organisms differ from one another (Reviewed in Wray 2007), and that individuals in a species differ from each other, including with respect to disease (reviewed in Mathelier *et al* 2015). Furthermore, these regulatory differences remain far more poorly understood than coding differences, and provide a source of perturbations by which the regulatory genome can be understood.

During my thesis I have used the model organism *Drosophila melanogaster* as a tool to analyze the regulatory genome. I have examined the transcriptional output of the developing *Drosophila* embryo, and the sequence motifs that control it, and have used information about genetic variation in gene expression to gain insight into the regulatory code.

I will begin by discussing current knowledge of the transcriptome and its regulation, its variation between tissues and cell types, and between individuals. I will touch on each of the data types and technologies used in my thesis work. In chapter two of my thesis, I describe my annotation of transcription start sites, (TSS) and 3' untranslated regions (UTRs) in *D. melanogaster,* my aggregation of chromatin immunoprecipitation (ChIP) and sequence motif data in *D. melanogaster* . Also discussed is a pipeline I used to select high quality transcription factor motifs, and an analysis of motifs present in polyadenylation sites (pA site) and TSS. In chapter three of my thesis, I describe my analysis of expression quantitative trait loci (QTL) during *D. melanogaster* embryonic development, with a particular view towards their genomic location, and the mechanisms by which they affect gene expression.

1.2 The Transcriptome

1.2.1 Methods in transcriptomics

Along with a new ability to measure genotype, the 21st century has seen a no less revolutionary shift in our ability to measure various molecular phenotypes. The first of these came with the development of DNA microarrays. By measuring the hybridization of fluorescently labeled DNA to complementary probes fixed to a solid substrate, DNA microarrays allowed the first high-throughput measures of gene expression. The technology (Schena 1995) only became practical with the advent of genome sequencing, since only known sequences can be queried. Microarrays have several limitations over sequencing-based technologies. Given issues with background hybridization signal, it is difficult to study low abundance transcripts with microarrays or to measure absolute, rather than relative, gene expression levels. They are also very limited (even in whole genome tiling arrays) in their ability to detect novel transcript isoforms. Microarrays remain useful for their cost efficiency and speed, particularly in genotyping applications where linkage increases the information they provide, but in transcriptomics research they have largely been superseded by RNA sequencing (reviewed in Allison *et al* 2006 & Shendure *et al* 2008).

As a more powerful means of measuring gene expression, RNA sequencing was initially practical only as a means to identify transcripts (e.g. Neto *et al* 2000), however next generation sequencing has now made it practical as a means to quantify RNA. The principle behind RNA-seq is simple; RNA is typically converted to cDNA, and after some intermediate steps like amplification and fragmentation, is sequenced. The resulting data consists of generally short (depending on the technology) reads whose count is proportional to the RNA's abundance in the original sample, as well as its length.

Because phenotypic information, unlike genotypic information, varies between cell types – in contrast to the near-uniform genome seen in different tissues of an organism - there is no single 'transcriptome'. Recently, the Encode

(Boyle *et al* 2014) and modEncode (Graveley *et al* 2011) projects have attempted to quantify gene expression levels in multiple developmental stages or cell lines. Increasingly sensitive methods have also yielded increasingly complex pictures of the transcriptome. Of particular note for my work are the various refinements that have been made to the RNA-seq protocol, which specifically focus on the 5' or 3' ends of transcribed RNAs.

Similar in essence to RNA-seq, CAGE (Shiraki *et al* 2003) uses chemical biotinylation of the diol group found on the 5' cap of eukaryotic mRNAs, to isolate capped transcripts. It then cleaves cDNA transcribed from these transcripts into short (27nt) 'tags', which correspond to the exact start site of expressed transcripts. By focusing sequencing on the 5' region of transcribed RNAs, CAGE allows single base pair resolution of transcription start sites.

Another refinement to the RNA-seq protocol is 'GRO-seq' a genome-wide application of the nuclear run-on technique, which isolates nascent RNA from transcriptionally competent nuclei by extending them with labeled nucleotides (Core *et al* 2008). The GROseq protocol suffers from the limitation that it gives only a 30-50bp resolution, which lead Kwak *et al* (2013) to develop PRO-seq, which uses single biotin labeled nucleotides to arrest the progress of transcribing polymerase, and thereby achieves base pair resolution. PRO-cap (or GRO-cap), a modification of this assay, uses a sequential 5' end dependent enzymatic treatment that ligates adaptors to capped transcripts, to select only the 5' start site of nascent transcripts. PRO-cap can thus be characterized as 'nascent CAGE ', and provides a measure of the level of transcript production at a site, rather than the steady state level of transcripts in the cell.

3'-Tag-Seq (Yoon and Brem *et al* 2010, Wilkening *et al* 2013) is another refinement of RNA-seq that like CAGE, yields single reads tagging individual transcripts, but at their 3' rather than their 5' end. The reads are obtained by fragmenting RNA and then amplifying it using primers against polyA, meaning that only polyadenlyated transcripts are tagged. This method is also necessarily somewhat less precise than CAGE in defining the location of transcript ends, since the end of a genomic polyA tract and the beginning of the added polyA tail cannot be distinguished.

3’-Tag-seq, CAGE and other tag-based methods allow for somewhat easier quantification of isoform levels, in that the proportion of tags at a site is proportional only to the level of the transcript, and not to the length of its various isoforms. Both techniques are insensitive to internal splice variants. They also suffer (particularly CAGE given the shorter length of the tags) from decreased ability to map transcripts, something that must be accounted for in studies of variation, since variation in genotype can lead to variation in capability, and hence apparent genotype-phenotype associations.

1.2.2 Understanding transcriptional complexity

RNA-seq and related methods have revealed a wealth of previously undocumented transcriptional complexity in both human (Mortazavi *et al* 2008) and model organisms (Cherbas *et al* 2011). Not only the pervasive transcription of eukaryotic genomes, but also the transcriptional complexity within individual genes, has become apparent. With alternative splice site, start site, and polyadenylation site usage being far more common than previously suspected (e.g. Carninci *et al* 2006 & Pelechano *et al* 2013), transcriptomics now faces the formidable task of determining what, if any, are the functional consequences. Isoforms that differ by their use of coding exons are, obviously, functionally distinct. However where transcripts differ only by their 5’ or 3’ untranslated regions – i.e. by the use of different TSS or pA sites – there may still be important functional differences, and these differences have only recently come to light.

Most eukaryotic mRNAs possess a 3’ “polyA tail” a string of adenines that are not templated by the genome but are rather added to the terminating transcript as it is cleaved and released from the transcribing polymerase complex (reviewed in Proudfoot 2011). The length of this polyA tail is one feature determining transcript stability. The 3’ UTR (and sometimes coding sequence) of a transcript can also contain sequence elements influencing translation, degradation, localization and termination (Proudfoot 2011). This means that alternative polyadenylation site usage can serve as a mechanism of gene regulation. Indeed, such processes are known to be crucial to the development of the *Drosophila* embryo (Semotok and

Lipshitz 2007). Factors regulating gene expression at the level of the mRNA include Pumilio, which stabilizes mRNAs encoding regulators of neurogenesis in *Drosophila* (Burow *et al* 2015) and Elav family proteins, (Dai *et al* 2011,Hilgers *et al* 2012). Elav, which is expressed in the developing nervous system of *D. melanogaster* binds to neural specific transcripts and suppresses the use of proximal polyadenylation cleavage sites. The resulting longer transcripts thus possess additional regulatory motifs, such as miRNA binding sites, allowing Elav to modulate neural gene expression by promoting alternative 3' isoform usage.

Differences in the 5' end of genes can also lead to functional differences. Many promoters make use of alternative TSS (ASS), resulting in transcripts using different 5' UTRs, and hence potentially different 5' regulatory motifs, upstream ORFs, or even different start codons (Pelechano *et al* 2013).

However even where differences at the 5' end are not themselves functional, they may point towards important differences in the underlying regulatory machinery generating the transcripts, as with the surprising degree of variation seen in TSS start site distribution (see e.g. Carninci *et al* 2006, Rach *et al* 2009, Hoskins *et al* 2011) using 5' CAGE and related methods. Single nucleotide resolution mapping of TSS starts has revealed that many genes have clusters of TSS rather than a single well-specified base pair at which transcripts are initiated. Promoters can thus be characterized by the spread of their TSS, which is often measured using some value analogous to the Shannon entropy of a probability distribution, (Rach *et al* 2009) such as the 'shape index' used by Hoskins *et al* (2011).

Such studies typically find that shape indices are not uniformly distributed, but rather are bimodaly distributed, and that genes can be loosely categorized as either 'broad' or 'narrow', with of course some promoters falling in between, and some studies additionally defining categories such as multimodal, or 'broad with peaks' (Rach *et al* 2009). These loose categories, furthermore, appear to define promoter classes with distinct regulatory architecture. Broad promoters are enriched for genes with housekeeping functions (Carninci *et al* 2006), which are expressed in most or all tissues. In contrast, narrow promoters tend to have more tightly focused expression patterns. Different 'core promoter' motifs are also known to be enriched in the two types of promoter, with for example TATA-box and

Initiator (INR) motifs being common in narrow promoters, and in *Drosophila* , DRE motifs being common in broad (Rach *et al* 2009 & Hoskins *et al* 2011). The two types of promoter also show different patterns of histone modifications (Kharchenko *et al* 2011), as well as preferences for distinct classes of regulatory element (Zabidi *et al* 2014). The exact means by which these two transcriptional programs differ remains obscure. It is likely to remain so until we better understand the functional properties of proximal regulatory elements like promoters and cleavage sites, and also the many other regulatory elements distributed throughout the genome.

1.3 *Cis* regulatory modules and their logic

1.3.1 *Cis* regulatory modules and sequence motifs

The term Cis Regulatory Module (CRM), encompasses any DNA sequence that acts to regulate gene expression in *cis* – i.e. when present on the same molecule of DNA. These include enhancers (which serve to promote transcription) silencers (which do the opposite), insulators, and also, promoters themselves. The terms denote functional activities that have been thought of as mutually exclusive sets of elements. However, in recent years it has become clear that a single piece of DNA can possess many activities. An enhancer may act to promote transcription in one context and suppress it in another (e.g. Herzog *et al* 2014). More puzzlingly, it has recently become clear that many segments of DNA thought of as enhancers act as TSS (albeit usually for low abundance, unstable 'eRNAs' – Kim *et al* 2011), and that the regions proximal to many gene's TSS can act as enhancers of other gene's transcription (Arnold *et al* 2013).

While the codon sequence of most protein coding genes is relatively easily identifiable from sequence alone (although with exceptions, such as very short ORFs), CRMs can be difficult to discern from non-functional 'junk DNA'. Naïve searches for transcription factor binding sites are wildly oversensitive – most individual transcription factors appear to be rather lacking sequence specificity in a genomic context, so that the activity of specific sequences arises in a nonlinear fashion from the interaction of many factors, including nucleosomes and their modifications (Zaret & Carroll., 2011), RNAs (Sigova *et al* 2015), and the biophysical properties of the DNA itself (Rohs *et al* 2009). Attempts to locate CRMs by sequence alone are thus limited by our knowledge of their underlying mechanisms.

Mechanistically, CRMs function as hubs for protein binding. The biochemistry of protein-DNA interaction has been the subject of much study. *In vitro* experiments, such as protein binding microarrays (PBMs – see Lee *et al* 2004) can be used to find the affinity of a factor for a large number of synthetic oligonucleotides. The results of such experiments can usually be summarized quite well by a Position Weight Matrix

(PWM). Though other more complex approaches do exist, incorporating non-independencies between base pairs, they have yet to see widespread use (reviewed in Boeva *et al* 2016). Many PWMs are not discovered using direct experimental methods like PBMs, but rather from sequence analysis, for instance from ChIP-seq experiments. A large number of tools exist, such as Meme (Bailey *et al* 2009), which attempt to find sequences enriched in a given set of sequences (such as binding sites identified with ChIP-seq). Methodologically they can be divided into methods like Meme, which try to learn the parameters of generative model, and enumerative models like Dreme, which search the space of consensus sequences for a motif maximizing some statistic. Such tools often output a large number of motifs. This is partially due to the actual biology of TF motifs, which tend to cluster with other motifs, and partially due to inevitable oversensitivity arising from the differences between approximate models of random sequence variation, and real mutational processes. Motif discovery tools are also, inevitably, underpowered compared to *in-vitro* binding assays, particularly in smaller genomes. The limited number of binding sites in the genome limits the information that can be gathered on a protein's binding preferences. There is thus a trade-off between the accurate, specific, yet artificial results of in vitro binding experiments, and the underpowered, nonspecific results of more natural *in vivo* PWM discovery methods. PWM databases are therefore of variable quality. Even good PWMs allow limited inference of actual protein occupancy. However, transcription factors do not bind as isolated units, but instead interact with other factors.

There is a good deal of evidence that the binding of factors at a locus is, in general, influenced by other factors, though this seems true for different enhancers and factors to varying degrees. The traditional model of transcription factor cooperativity (e.g. Frank *et al* 1990) where individual TF's physically interact, is one such mechanism, but other mechanisms exist by which TFs influence each other's binding, some more direct than others. 'The enhanceosome' (Panne *et al* 2008) is a classic example of extreme cooperatively. This well-studied cluster of transcription factors, which bind upstream of the Beta Globin locus, each protein has a protein-protein interaction with its neighbor. This is facilitated by the correct position (spacing, orientation) of each of the six or more transcription factors that occupy it.

Numerous other characterized physical interactions between regulatory proteins exist, such as the cooperativity between *Drosophila* TFs Dorsal and Twist in the developing embryo (Zinzen *et al* 2009).

In contrast, the binding of TFs at many loci appears to function in more of a 'billboard' fashion, with groups of co-binding factors occupying the enhancer independently of each other (Zarett 2011). The 'TF collective' model is somewhere in between with many TFs binding to the same enhancers, apparently showing non independent binding (since factors can bind even in the absence of their motifs, if other members of the collective are present), but without the requirement for a rigid motif grammar that would indicate direct physical interaction (Junion *et al* 2012).

In addition to direct protein-protein interactions, a number of mechanisms can mediate indirect cooperativity. One such mechanism is the steric hindrance of transcription factors by nucleosomes. Nucleosomes are known to prevent binding of transcription factors to DNA in biochemical assays, and while the situation in the cell is more complex, the strong correspondence of nucleosome-ChIP occupancy to DNase measured DNA accessibility strongly suggests that occupation by nucleosomes is a crucial factor regulating transcription factor binding in the cell as well. Certain factors, such as Vfl (Zld) in *Drosophila* (Sun *et al* 2015), and FoxA factors (Zaret & Carroll, 2011) in human hepatocytes, appear to act as 'pioneer factors', which facilitate binding by other 'settler' TFs by rendering binding sites accessible. A second mechanism by which indirect cooperativity may occur is via physical interaction with cofactors. TFs often exert their effects on gene regulation by recruiting cofactors, and given that such factors are often bound by multiple TFs (Stampfel *et al* 2016), they may act as a physical bridge between transcription factors.

The result of this mechanistic complexity is that while sequence motifs are a valuable tool for understanding the regulatory genome, they are not sufficient to identify CRMs, or more importantly for predicting their functional output and the effects that variation will have on this. For this, other tools are needed.

1.3.2 Detecting *Cis* regulatory modules genome wide

In general, the absence of clear sequence rules to identify 'all' CRMs in a genome has meant less direct methods become necessary. For many years, CRMs have been identified by functional assays. An enhancer, for instance, is typically defined as a sequence that can enhance transcription in a manner independent of its orientation when cloned upstream of a reporter gene (Banerji *et al* 1981). Such assays have been useful for studying specific regions, but their low throughput limits their application to the genome as a whole.

Fortunately, the genomics era has given us many new tools to identify CRMs' location, abundance and function. Comparative biology is the oldest means of identifying regulatory sequence. Such methods have been particularly effective in vertebrates, where they are facilitated by the large genome sizes of the model organisms (Peterson *et al* 2009). In principle, a sequence that encodes regulatory information, should be under purifying selection to maintain some sequence properties, and this purifying selection should be detectable both in the form of reduced variation between species, and reduced variation within species. Positive selection, while rarer and harder to detect than purifying selection, can also indicate regulatory sequence. In practice however, neither may be detectable, due to the turnover of functional sequence and the complex relationship between sequence and function, however comparative biology remains a valuable tool.

Natural selection acts on DNA sequence to both increase the frequency of adaptive variants (positive selection) and, more commonly since most mutations are neutral or deleterious, to decrease the frequency of deleterious substitutions (negative selection). Both of these phenomena leave characteristic signatures in the genome and a large number of methods exist to detect them, using either intra species, or inter species comparison. Conservation between species is an obvious indication of functional sequence under negative selection, and many tools exist to align and compare sequences in order to detect it e.g. Phastcons (Siepel 2005). At the intra-species level, an excess of low frequency alleles within a region (detected via e.g. Tajima's D) can also be an indication of selection (or sometimes population stratification etc. (Tajima 1989)). Other simple tests exist for positive selection, for

instance, the MacDonald-Kreitman test (MacDonald & Kreitman, 1991). This tests the hypothesis that a given genomic region (such as the non degenerate coding bases of a gene) shows a different ratio of between-species to within-species polymorphism, than a control set of regions (such as the 4d degenerate coding bases of a gene). Adaptive substitution, by driving polymorphisms to fixation, will generate divergence between species greater than that predicted by the within species variation alone.

In general, scans for selection over large timescales, using only interspecies comparisons, are not useful for the analysis of invertebrate regulatory regions; the turnover of functional sites makes accurate alignment too difficult (Hare *et al* 2008). Methods that focus on shorter timescales and use intra species variation are therefore more suitable. Another difficulty is that since positive selection and negative selection will leave opposing marks on the genome, simple tests can give biased results where both are present in a set of regions. INSIGHT (Inference of Natural Selection from Interspersed Genomically coHerent elemenTs - Gronau *et al* 2013) is a recently developed method that integrates between and within species variation data to estimate rates of both positive and negative selection simultaneously, using a maximum likelihood based approach to fit a model incorporating both as parameters. INSIGHT pools information from many short, interspersed regions. It also accounts for differing mutation rates across genomic regions by estimating divergence and polymorphism locally, using control regions that are assumed to be neutral, similar to the use of control regions in a MacDonald Kreitman test. INSIGHT's results reflect recent levels of natural selection, and are therefore robust to alignment problems caused by turnover of functional sights. During my thesis, I make use of INSIGHT to analyze intra species variation in *D. melanogaster* (using inbred lines from the *Drosophila* Genetic Reference Panel) using data on inter-species variation for ancestral sequence inference (Clark *et al* 2007). The INSIGHT framework allows detection of selective effects in interspersed genomic elements like promoters and sequence motifs, which are not suitable for analysis with simpler methods such as the McDonald-Kreitman test.

In practice however, comparative sequence analysis is a highly imperfect method of detecting regulatory sequence, for several reasons. The first is that all

methods of detecting conservation have imperfect sensitivity; regulatory sequences may evolve rapidly despite sequence constraints. This can occur if there is a large space of neutral sequences through which they can move, or, as seems to be the case, if positive selection frequently causes them to diverge from one another (Hare *et al* 2008 & He *et al* 2011). Relying on genome sequence alone then, can only ever give a broad picture of the regulatory genome, though nevertheless a valuable one, since the approach is unbiased towards any particular tissue or assay.

Luckily, particular assays for examining the regulatory genome are not in short supply. One example is STARR-seq, an experimental technique that essentially caries out many functional enhancer assays in parallel, by having each enhancer transcribe itself, thereby providing a multiplexed read out of each enhancers activity. Although it provides a direct measure of enhancer activity, STARR-seq (like traditional reporter assays) measures an enhancers activity outside of its normal genomic context, and may not, therefore be perfect guide to actual regulatory activity. STARR-seq is also applicable only to cell culture, relying as it does on the transfection of a large library of enhancers.

Other assays rely on a number of biochemical features common to regulatory sequences that possess 'enhancer activity' as per traditional assays. Like transcribed regions, they are typically more exposed to enzymes and other proteins than non-functional sequences. Several assays such as 'DNase-seq' (Thomas *et al* 2011) take advantage of this property to generate reads in proportion to a sequence's accessibility, and thereby create maps of chromatin accessibility (although other factors – such as DNA conformation and the sequence biases of the assay, affect the results). The information from accessibility based methods provides a broad view of which genomic regions are likely to be active in a given sample, but it does not, alone, give any information about *which* factors are likely to be bound at the region. The combination of specific accessibility patterns and underlying sequence information can, for some factors and organisms, give good probabilistic information about the factors at a region (Pique-Regi *et al* 2011). However, not all factors have distinctive enough sequence preferences and accessibility patterns for footprinting to be effective. Furthermore, footprinting is necessarily under-sensitive, even in the best circumstances because not all bound regions for a factor will leave identifiable

footprints. For more sensitive survey of occupancy, ChIP based methods are required.

ChIP-chip and ChIP-seq both make use of antibodies specific against a factor of interest to isolate regions of DNA bound by the factor. Chromatin Immunoprecipitation is a well-established technique first used in 1988 to examine DNA-histone interactions (Solomon *et al* 1988). With the advent of microarray and sequencing technology however, it has become a high throughput means of interogating genomic protein occupancy. The technique has therefore taken on a prominent role in the field of functional genomics, with ChIP experiments and their variations taking a prominent role, for instance, in the ENCODE and MODENCODE projects. Both ChIP-chip and ChIP-seq involve the immunoprecipitation of proteins to bound DNA, generally using a fixation agent like formaldehyde, and both yield similar results; a quantitative measure of a factor's occupancy throughout the genome, from which regions enriched over background signal (peaks) are called. The techniques differ primarily in their resolution – with the regions called by ChIP-seq often being an order of magnitude smaller, since they are not limited by the tiling density of genomic probes. As with RNA-based technologies, a variety of refinements to the basic assay exist, which vary for instance, the choice of fixation agent (Zhang *et al* 2004), or use of enzymatic digestion to give very precise measurements of binding position (Rhee & Pugh, 2011).

The plethora of ChIP-seq data now available contains occupancy data for a huge variety of factors. DNA-binding factors like TFs and insulator proteins, DNA-polymerases and isomerases, and histone variant and post-translational modifications. Maps of histone mofidifications in particular have been instrumental in our understanding of the regulatory genome, with characteristic patterns (chromatin signatures) appearing to define different functional classes of region (e.g. Kharchenko *et al* 2011, Ho *et al* 2014). Enhancers and promoters are charaterized by very similiar patterns of chromatin marks, with H3K4me1 and H3K4me3 usually being present at both, and with a higher level of the latter at promoters. Many papers have made use of chromatin marks and other such genomic signals to create operational definitions of enhancers. It is crucial to remember, however, that biochemical marker characteristics may not correspond neatly to actual function.

Indeed chromatin immunoprecipitation data is is often a relatively poor guide to function. One surprising result that has consistently emerged from ChIP experiments is that transcription factors occupy a very large number of sites in the genome. Functional experiments such as knock-downs (Cusonovich *et al* 2014), as well as sequence conservation (Andolfatto 2005), strongly suggest that the vast majority of such sites are nonfunctional. In part, this may be because the 'occupancy' measured by ChIP experiments masks important biochemical differences in the occupancy at different sites. For instance, studies that use competition between two labelled transcription factor constructs have shown that ChIP occupancy measures only the time spent bound to DNA, irrespective of the duration of each occupancy event (Lickwar *et al* 2012), and therefore includes a large number of sites which are bound only transiently. These 'treadmilling' sites appear to be less functional, on average, than sites with less transcient occupancy, which may account for some of the 'surplus' occupancy in most genomes. It is likely however that a great deal of occupancy is of little functional consequence because, as discussed, transcription factors function combinatorially, rather than in isolation. It thus becomes critical to analyze the regulatory function of DNA directly, either by pertubative experiments, or by taking advantage of the natural perturbations caused by genetic variation.

1.3.3 *Cis* regulatory modules as transcriptional elements

In 1992, the Beta Globin locus control region (LCR), an enhancer, was shown to be transcribed. The discovery remained something of a curiosity for almost 20 years, until Kim *et al* (2010), observed that DNA Polymerase II bound a large number of enhancers in neurons and actively transcribed them. Since this study was published, a large number of subsequent studies (e.g. Andersson *et al* 2014 Kim *et al* 2010, Wu *et al* 2014) have documented the presence of eRNAs in mammalian systems. eRNAs are typically low abundance, non-polyadenlyated transcripts that are associated with the activity of the enhancer. With only a few papers mentioning its presence (Karchenko *et al* 2011) and greater directional bias than in mammals (Core *et al* 2012), enhancer transcription has received less attention in *Drosophila* .

One obvious explanation for the of enhancer transcription is that, because eukaryotic transcriptional machinery is relatively nonspecific (Johnson *et al* 2005), and because active enhancers have several transcription-promoting properties including open chromatin and high concentration of transcription factors and polymerase (e.g. Negre *et al* 2011, Kharchenko *et al* 2011, Thomas *et al* 2011), transcription of eRNAs represents a non-functional by-product of enhancers' normal activity, which is to recruit transcription factors to active promoters. The degree to which eRNAs are produced in an enhancer would then be correlated with its activity, but would also depend on sequence features of the enhancer i.e. enhancers which were incidentally 'promoter-like' would tend to produce eRNA.

Another, related theory is that while eRNA itself is non functional, its production results in the functional recruitment of polymerase and other factors, whose presence do affect enhancer function. This theory predicts that sequence features associated with its production will be conserved to some degree, although their position may not be. It also predicts that disrupting eRNA production after transcription will not affect enhancer function but disrupting its actual production will. Distinguishing between these two theories thus requires precise perturbations of the molecular chain of events that occur at enhancers.

Both of the above scenarios however predict the presence of nonfunctional transcripts around functioning promoters. Under such conditions, we should not be surprised if eRNA becomes a substrate for the evolution of enhancers. With RNA present and capable of interaction with DBFs and their cofactors, selection for increased binding at a locus should favor mutations in eRNA that increase affinity, just as it favors mutations in the DNA that increase affinity. In support of this, a good deal of anecdotal evidence now exists that enhancer transcription is in some cases functional (e.g. Mousavi *et al* 2013,Sigova *et al* 2015, Schaukowich *et al* 2014). However such studies are inherently biased, in that they will tend to focus on eRNAs with particular properties such as higher expression levels and proximity to well studied genes. It remains unclear what proportion of eRNAs are in fact functional even in mammalian systems, and still less clear what proportion, if any, of *Drosophila* 'eRNAs' might be functional.

There is no a priori reason to think that eRNAs share functional characteristics, given that they do not, in general, share an evolutionary origin (in contrast to for instance a family of proteins). The current state of the evidence would appear to suggest that all of the above theories are functional for at least some eRNA. In this case then, the task becomes to determine not if eRNA *can* be functional, but rather how often they are, and how.

Recent studies have addressed the tissue specific nature of eRNA expression. Regardless of their biological function, eRNAs have proven to be a good indicator of regulatory activity (Yao $et\ al$ 2015, Andersson $et\ al$ 2014,Wu $et\ al$ 2014). While eRNAs seem to be present at a baseline level throughout development, they also appear to show increased expression in the tissues in which their enhancers promote expression, suggesting a quantitative relationship between enhancer activity and eRNA expression. The nature of this relationship however, remains unclear.

As a model system, *Drosophila* potentially offers a number of advantages for studying eRNAs including; a wealth of existing data, annotated regulatory elements, and the use of vivo expression assays to investigate eRNA function in the context of living tissues.

1.4 Genetic variation

1.4.1 Neutral evolution, and genetic linkage

It was to explain the patterns of variation between individuals that Mendel originally devised his genetic theory of inheritance. During the 20th century, genetics developed into an experimental, molecular science, which led to a separation between population level genetics and molecular genetics, fields that are now being reconnected. Ronald A Fisher was the first to show that Mendel's discrete units of heredity (genes) could give rise to the smooth, quantitative patterns of variation seen in actual populations (Fisher 1921). Mid 20th century explorations of genetic variability, such as observations of allozymes and restriction fragment length polymorphisms revealed a surprisingly large degree of variability between organisms. It soon became clear that a great deal of the genetic variation between organisms must be neutral or effectively so in its effects, with interspecies and intra-species variation being in large part a result of "genetic drift", rather than evolution under natural selection.

Motoo Kimura (1968) and others authored theories of neutral evolution to reflect this new reality. With the advent of genome sequencing, the full extent of variability between organisms became clear – any given pair of humans – even closely related ones – differ in thousands of genomic locations, both by single nucleotide polymorphisms (SNPs) and by larger structural variants such as deletions, inversions and insertions. These differences are even greater in other species such as *D. melanogaster* (Mackay *et al* 2012), which have much greater effective population sizes.

Functional variants are thus needles lost in the haystack of neutral variation. The overwhelming number of differences between conspecific obviously represents a challenge to any attempt to link phenotype and genotype via statistical association alone. Moreover, because genetic variants are physically linked to one another by their location on the chromosome, they are not inherited together, but instead are in 'linkage disequilibrium' (LD) to one another in (nonlinear) proportion to their

chromosomal separation. In humans, recent population bottlenecks and the nonrandom frequency of recombination in the genome means that variants travel together in large 'LD blocks'. By contrast, in other organisms like *Drosophila* linkage is far less extensive (Mackay *et al* 2012).

When mapping the approximate location of variants responsible for a particular trait, or detecting natural selection, linkage disequilibrium is an advantage. Early studies of monogenic disease in humans for instance were able to use easily detectable natural variants as markers to map genes for disease like Huntington's chorea (Gusella *et al* 1983), and soft selective sweeps (Garud *et al* 2015) can also be detected as a result of LD. Linkage also allows genotyping arrays covering a relatively small proportion of SNPs to mark most of the common variation in the genome, which has allowed early genomics projects like the HapMap (Frazer *et al* 2007) project to survey common genetic variation with microarray genotyping at a fraction of the cost that would otherwise be required.

1.4.2 Expression quantitative trait loci

In the years since the human genome project, countless genome wide-association studies have been performed. Genome wide association studies, and the closely related eQTL studies, are in essence simply a modern extension of the simple genetic association studies which have in use for decades, for instance in agriculture. Individuals with a genetic variant are compared to those without, and a statistical test, such as a linear regression (for Quantitative Trait Analysis) or a Chi-squared test (for binary traits) is applied to test the null hypothesis of no association. If the p-value passes the significance threshold, an associated locus (or Quantitative Trait Locus for quantitative traits) has been identified.

eQTL studies differ from traditional quantitative trait analysis in that individual gene expression values are used as the quantitative trait, and that the whole genome can now potentially be interrogated for associations. When testing thousands of variants however, corrections for multiple hypothesis testing (or a Bayesian framework with a conservative prior) must be applied such as the use of false positive rates as summary statistics, and stringent thresholds. This means that genome wide searches inevitably suffer a loss of power, particularly eQTL studies,

where each gene is also interrogated. Analyses are thus usually only carried out on a subset of variants close the gene in question. Typically the area chosen is a *cis* window around the gene ~ 200kb in size, an empirical decision that has since been justified by the finding that *cis* regulatory interactions are for the most part confined to topological domains (TADs) of approximately this size (Ghavi-Helm *et al* 2014).

False associations due to population structure can also lead both GWAS and eQTL studies astray. Modern techniques (e.g. Lippert *et al* 2014) control for population structure by various means, such as the use of mixed linear models incorporating covariance structures based on the genetic distance between individuals.

Both GWAS and eQTL studies have also reached a turning point in recent years. With countless eQTL and GWAS hits identified for various traits and tissues, it has become clear that translating these associations into biological understanding is less than trivial. Associations in humans can typically only identify large blocks of linked variants, and thus, except when a variant in coding sequence is present that clearly disrupts protein function, the exact identity of the causal variant, and its mechanism, is often unclear. Furthermore in both classes of study, efforts have begun to analyze groups of studies (Flutre *et al* 2013) in an effort to understand the causes of the relatively low reproducibility between them. Some of this lack of reproducibility has been explained using more sophisticated joint analysis methods, however a good deal of it seems to be genuine heterogeneity between tissues and genetic backgrounds (Huang *et al* 2013).

Recent studies (Gaffney *et al* 2012, Das *et al* 2016) have attempted to leverage other genomic data sets such as motif, ChIP and DNase data, to make more accurate inferences of causal variants. Such studies not only yield better information on the causal variants, but also, insights into the relative importance of different genomic elements in genetic regulation. Such studies have shown that signatures associated with transcribed regions, active enhancers, and TSS are all statistically enriched for eQTL. As methods for predicting causal variants however, they necessarily introduce a bias towards known, studied genomic features. As such, they run the risk of biasing studies away from the most interesting variants – those with novel, unstudied mechanisms.

During my PhD I have used several QTL datasets called in *D. melanogaster* to analyze the distribution and mechanism of variants affecting gene expression in the developing fly embryo. Working in *Drosophila* provides access to datasets which are unavailable in human, such as panels of regulatory regions which have had their spatiotemporal regulatory properties assayed using a controlled vocabulary (Kvon *et al* 2014), and the chance to validate observed QTLs using whole embryo expression data. Moreover, the *D. melanogaster* genome shows far less linkage disequilibrium than mammalian systems; where human or outbred mouse populations have linkage blocks of 20kb to 100kb (Cannavo 11,12), *Drosophila* typically shows blocks of less than 5kb. Moreover, the *Drosophila* genome is highly variable, with approximately one SNP every 100bp (Mackay *et al* 2012). These traits combine to allow very high resolution QTL mapping in *D. melanogaster*.

Drosophila also the provides opportunity to make use of data from controlled breeding schemes like the *Drosophila* Genetic Reference Panel (DGRP), which would of course be unethical in humans. The DGRP (Mackay *et al* 2012) is a panel of *D. melanogaster* lines, which have been inbred to near total homozygosity. As such, they represent a sample of wild type variation from the *D. melanogaster* population, and can be used in eQTL studies in an analogous way to panels of human cell lines such as the HapMap cell lines. Studying genetic variation in the DGRP offers several unique opportunities. Cell lines cannot be studied at varying points in development in a way that accurately recapitulates developmental processes. Furthermore, the homozygous genotypes of these lines will allow increased power to detect effects on gene expression, since variance between lines is increased, and recessive and co-dominant effects are rendered stronger and more detectable than in heterozygous individuals.

In the presented work I have leveraged these features of *Drosophila* as a model system to gain insights into genetic variation in gene expression, making use of both high-resolution QTL data and *in vivo* expression datasets to study the mechanisms of variation in gene expression.

RESULTS

2 The Transcriptional Landscape of the Developing *D. melanogaster* Embryo

The complex sequence of events that underlie the development of a D.melanogaster embryo from a single cell to a multicellular, motile larva is coordinated by a complex array of transcripts. Produced over early development or inherited from the transcriptome, these transcriptional events reflect the regulatory cascade that shapes the undifferentiated syncytium's genome into that of the many differentiated cell types.

During my PhD, I have studied the transcriptional landscape of the developing fly embryo using two high throughput gene expression datasets. Both datasets were collected in developing embryos at three time points, in ~80 different genotypes (using a panel of inbred lines, as part of the DGRP consortium – Mackay *et al* 2012). One, generated using 5' CAGE, was used for an eQTL analysis of 5' start sites, while the other, generated using 3' Tag-seq, was used in an eQTL analysis of polyadenylation sites. In this chapter I discuss my efforts to annotate and characterize TSS using the 5' CAGE data, and to characterize 3' UTR regions using the 3' data. These efforts provided datasets for use in later analyses in the two projects, as well as information about *D. melanogaster* biology in their own right. Here I discuss their implications for the biology of *D. melanogaster* transcriptomics and development.

2.1 An in depth characterization of transcription start sites in *D. melanogaster*

In order to analyze the causes of variation in transcription start site usage in developing *Drosophila* embryos, a collaborator in the Furlong lab collected 5' CAGE data from embryos in 82 DGRP lines at three time windows of embryogenesis, in whole embryos. RNA was isolated from embryos at 2-4 hours (stages 5-8), during

which cells remain multipotent, 6-8 hours (stages 10-11), during which the major cell lineages are specified, and 10-12 hours, during which terminal differentiation is occurring (stages 13-14). To annotate the location of TSS's, we reasoned that differences between lines should be of a quantitative, rather than qualitative, nature, and so pooled the data for each line together to create a single dataset for each time point, at unprecedented sequencing depth at these stages of embryogenesis. Since the DGRP lines represent the variation in the wild type population, this dataset should reflect the use of TSS in the population as a whole.

The dataset, containing over 2.5 billion reads, represents a substantial improvement in both time resolution and sequencing depth compared to previous *D. melanogaster* start site annotations using CAGE (Hoskins *et al* 2011). We first attempted to make use of the hierarchical clustering based approach used by Hoskins *et al* to define more biologically meaningful TSS. However this approach failed badly due to the high depth of our data – given that some of our most expressed genes have >50% of their length tagged, it becomes impractical to tag the TSS without in some way taking their magnitude into account. I therefore made use of my own algorithm to identify a set of high quality TSS.

2.1.1 Modeling noise in highly expressed genes

One prominent feature of our data that made identifying TSS challenging was the degree of noise seen at highly expressed genes. Visualization suggested that most highly expressed genes showed a baseline level of CAGE reads along their length, which is likely, a result of both transcript re-capping, and contamination from uncapped RNA. Previous studies have attempted to compensate for this phenomenon by assuming that it followed the same distribution as RNAseq data for the same tissues (Hoskins *et al* 2011). We found that published RNAseq datasets did not offer sufficient coverage over many genes to model the noise effectively, and so we elected to model the noise as being flatly distributed over gene bodies, with its magnitude being directly proportional to the number of CAGE reads over a gene. We reasoned that this assumption would be a reasonable one, since CAGE reads should be roughly proportional to RNAseq signal. We then modeled the data as being a mixture of this flatly distributed noise, whose distribution was represented by a beta

binomial distribution, and our signal (see materials and methods). We chose the beta binomial distribution in accordance with the method used by Hoskins et al. A binomial distribution is suitable for describing the results of a stochastic process in which discrete events (reads) are assigned to a particular outcome (a given site), with some unknown probability. Modeling uncertainty about the probability of the individual events with a beta distribution allows the total uncertainty about the process to be modeled by a beta-binomial distribution. I used a simulated uniform distribution in much the same way that Hoskins $et\ al$ used RNAseq data as prior information about the probability parameter for each site (Fig 2.1).

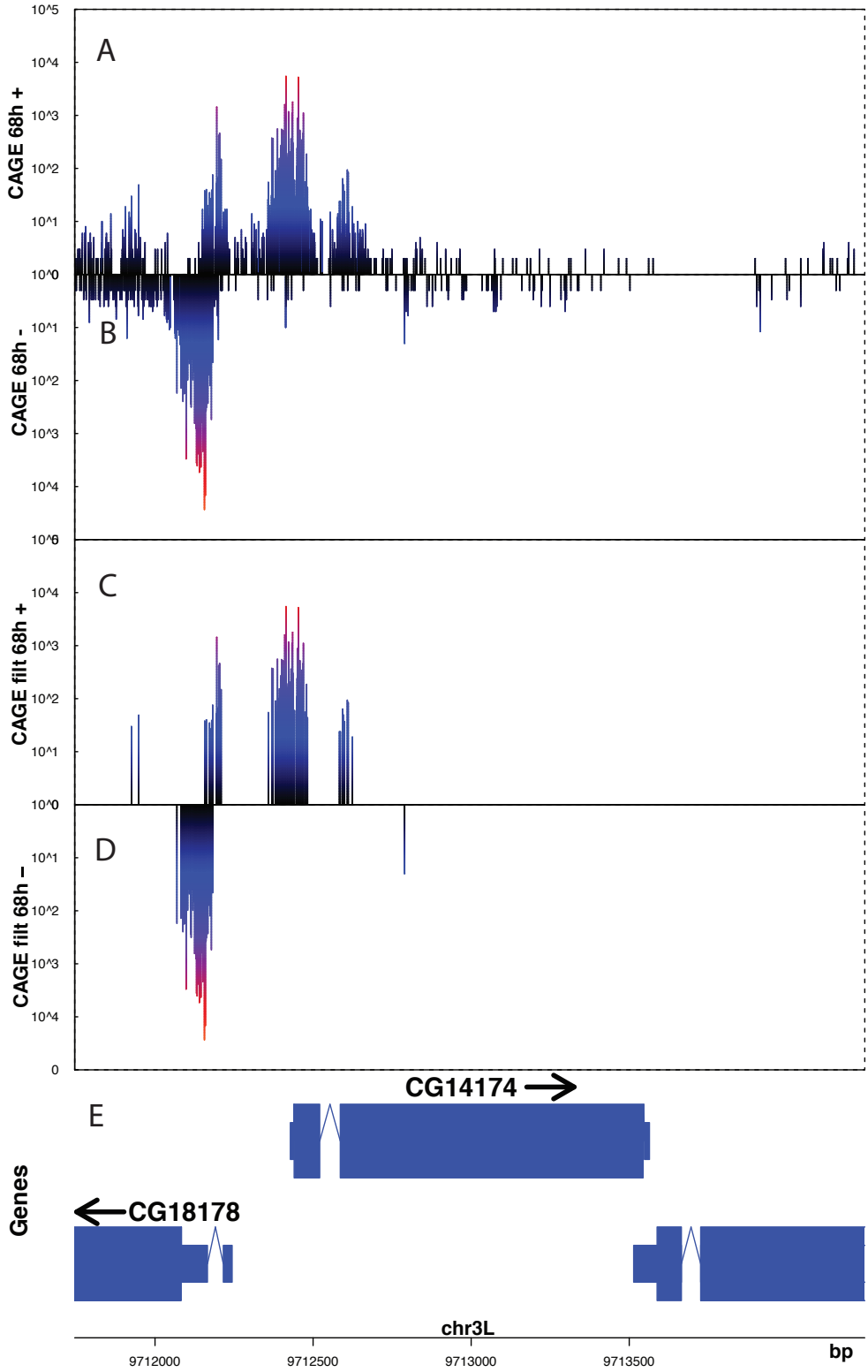


Figure 2.1: Uniform noise filter for the CAGE data.

From top to bottom, panels represent: A) Raw CAGE tag density for the positive strand over the gene *CG14174,* plotted on a log10 scale. B) Same, but for the negative strand C) The CAGE tag density after application of the uniform noise filter

to the data, positive strand, D) Same, for negative E) Gene model for *CG14174* and surrounding genes, arrows indicate gene orientation. Note that some CAGE tag peaks remain internal to the gene after noise filtering, due to the heterogeneous distribution of the noise.

2.1.2 Defining peaks using local smoothing

The next step in my peak calling procedure was to delineate regions of high CAGE expression. Most studies of enhancers define a 'core promoter' to be a region of between 100 and 300 bp in length (e.g. Ohler *et al* 2002). I experimented with various transformations of the data and found that a simple local averaging operation, with a bandwidth of 150 bp, yielded peaks that corresponded well visually to known transcript start sites and motifs. 94% of genes in the top 50% of the expression distribution at 6-8 hours, as assessed by the total read count over their gene body, have a cage peak in the 'main' set located over an annotated transcript site. Moreover, the central 10% of our cage peaks was 23-fold enriched for annotated TSS compared to the outer 10% ($p<10^{-16}$, Fisher's Exact test). By then taking local minima in this smoothed data, we were able to separate our CAGE signal into peaks for further filtering and analysis see Fig (2.2).

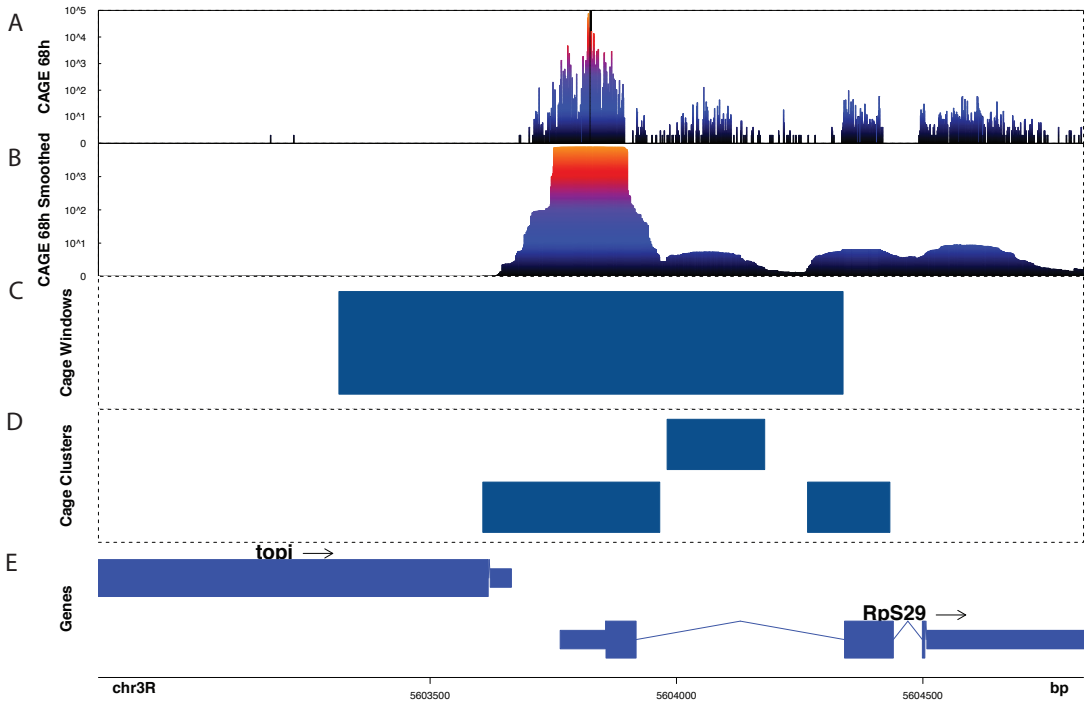


Figure 2.2: Defining CAGE peaks using smoothing.

From top to bottom, panels represent: A) Filtered CAGE tag density for the positive strand over the gene *RpS29,* plotted on a log10 scale, for positive strand only. B) The above, smoothed using a moving average over a 300bp window. C) 1kb CAGE tag window, centered on point of highest expression for the gene D) CAGE peaks, defined as boundaries between the local minima in the smoothed CAGE tag distribution, and filtered for proximity to annotated start sites. E) Gene model for *RpS29* and surrounding genes, arrows indicate gene orientation.

We found that due to the irregular structure of noise within gene bodies (perhaps caused by exon structure, splice junctions, re-capping sites etc.) we were often left with large numbers of peaks internal to our genes. We therefore decided to further sub-select our peaks by attributing them to genes, and selecting only peaks likely to represent annotated transcripts, and which accounted for more than 5% of the gene's total expression (see Materials and methods for details).

2.1.3 Peak characteristics – number, shape, associated genes

Having identified high-confidence TSS peaks, I then carried out some analysis of their location, and, broadness, and number Fig (2.3). Our peak set contains 97% of the total CAGE signal, with 95% of this falling within the 'main' set. Our 'main' set of peaks are distinguished from 'internal' peaks by several characteristics – internal peaks are of course less expressed (unsurprising given the filtering process), with the main peaks having a median tag count of 16809 across all three timepoints, compared to the internal peaks' median of 576 (almost 30 fold higher). Internal peaks tend to be found on genes with higher overall expression.

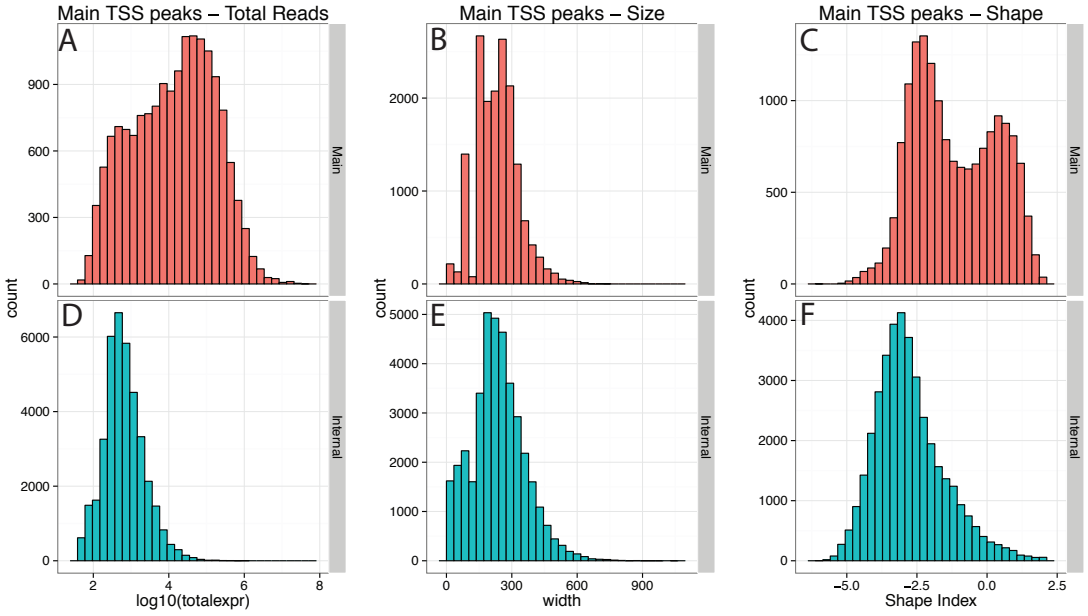


Figure 2.3: Main vs. Internal clusters.

Descriptive statistics for the 'Main' peak set – i.e. TSS proximal peaks accounting for a large fraction of gene's expression after noise filtering, and internal peaks, which occur within the gene body. A, D) Expression level in Log10 total CAGE tag counts (x-axis) summed over three time points. B, E) Size in bp for Main and Internal CAGE peaks(x-axis). C, F) Distribution of shape index - a metric of how focused transcription is at a particular site. Internal (blue, D-F) CAGE peaks are in general less expressed than the main (red, A-C) set, and in general have low shape indices, whereas the main set show a much higher range of tag counts, and are often very narrow.

Previous studies have often categorized peaks by their degree of broadness. I elected to use the descriptive statistic used by Hoskins *et al* – 'shape index' to characterize enhancer shape. The shape index is essentially the entropy of the distribution of tags over a promoter. It measures the degree to which reads or focused on specific base pairs of the promoter, and does so in a way independent of promoter width, which is influenced by the presence of rare outlier TSS and other noise, particularly in high-depth datasets such as ours. I found that both CAGE Windows and CAGE peaks showed a bimodal distribution of shape index (Fig 2.3c). The midpoint between the two modes of this distribution lies at approximately -1, and so we refer to peaks with a shape index of -1 or greater as 'narrow' in subsequent analyses. The lower (broader) end of the shape distribution spectrum is heavily biased towards internal peaks. This likely reflects the fact that 'internal' peaks tend to occur in transcribed regions where the distribution of tags tends to follow a much flatter distribution than it does in promoters. In contrast to some previous reports (Hoskins *et al* 2011) I find a clear bimodality in the distribution of shape index for *Drosophila* promoters. We also carried out gene set enrichment analysis (GSEA) to assess which genes had greater numbers of CAGE peaks in the main set. Since both expression and gene length likely co-vary with the number of peaks, we controlled for these by fitting a negative binomial, with the number of main peaks for a gene as the response variable, and it's length and expression as predictors. The residuals from this model were then used to carry out GSEA. Genes with larger numbers of peaks in the 'main' set are enriched for GO biological process terms associated with development and signal transduction, illustrating that genes in these categories tend to have more complex architecture (Fig 2.4). Unexpectedly, certain metabolic processes, such as lipid metabolism and peroxisomal transport, are also enriched for high peak numbers.

Previous work has shown that certain motifs are associated with broad and narrow promoter shapes (e.g. Hoskins *et al* 2011). I could confirm that indeed, motifs such

as the TATA-box, INR, and DPE showed a tendency to be found in narrow promoters, while motifs like DRE and Motif 1 were more common in broad peaks (Fig 2.5).

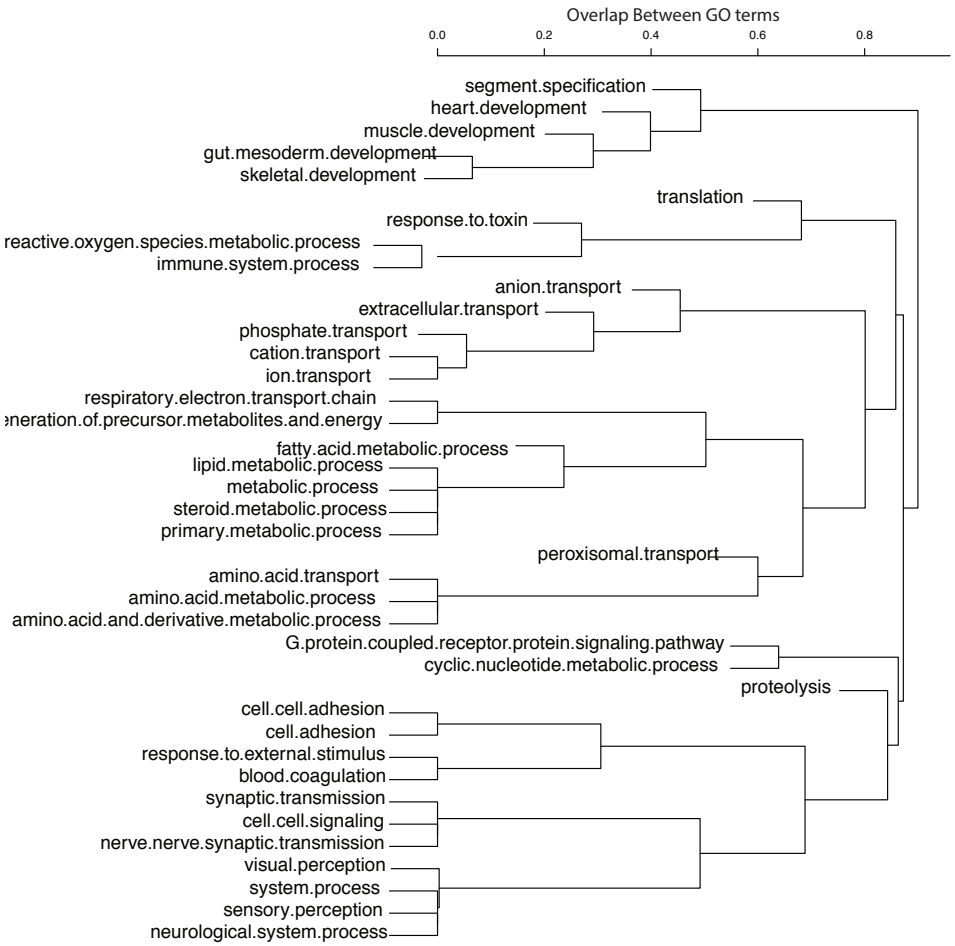


Figure 2.4: Go terms associated with genes with greater numbers of TSS.

GOslim biological function terms associated with genes with larger numbers of TSS in the Main set of CAGE peaks. The number of 'main' CAGE peaks associated with each gene was counted, and a negative binomial regression was used to control for correlation with gene expression and gene length. GOslim terms with 50 or more genes enriched for peak number were plotted ($p<0.05$, point biserial correlation - see materials and methods). Clustergrams reflect the fraction of genes shared by any two categories, with 0 indicating nested categories and 1 reflecting orthogonality.

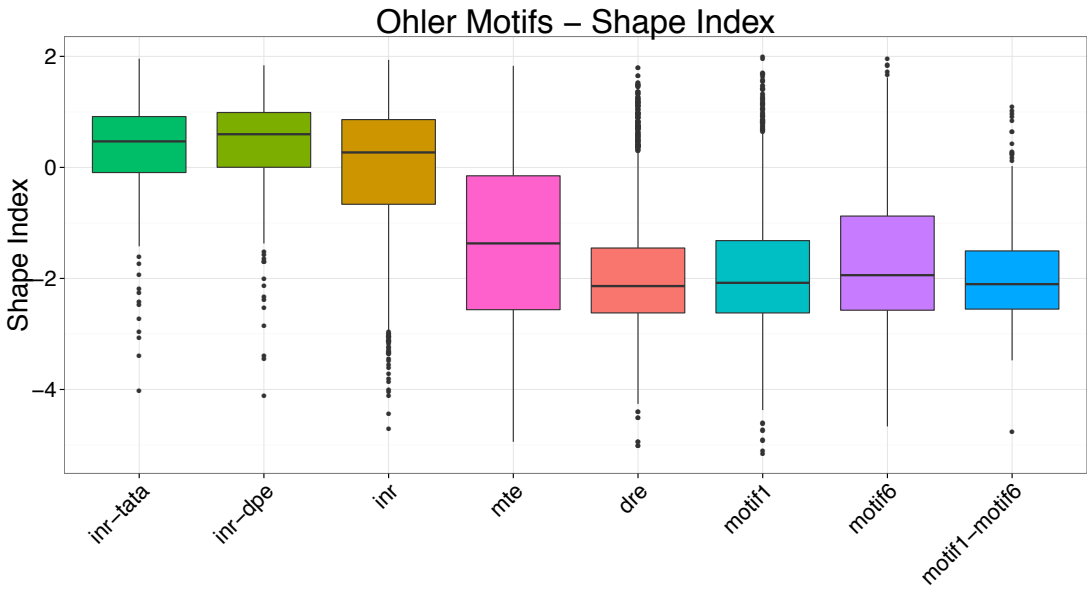


Figure 2.5: Association between promoter motifs and shape index.

The distribution of shape index (y-axis) for CAGE peaks with characterized promoter motifs. Shape index (SI) measures how clustered tags are within a promoter, with an SI of 2 indicating all tags are at a single basepair. INR, the TATA box, and DPE are associated with narrow clusters, as expected, while the DRE motif and *Drosophila* promoter element 1 and six show association with broader promoters, with MTE showing an intermediate distribution.

2.1.4 CAGE peaks – differential expression during development

Previous work has shown (Hoskins *et al* 2011, Carninci *et al* 2006) narrower promoters tend to show more dynamic, specific transcriptional patterns. I wished to confirm that this was true of my CAGE peaks, and to do so, I used the DESeq2 package (Love *et al* 2014) to test peaks for differential expression between time points, and examine which promoters showed changes in expression over the course of *D. melanogaster* development (Fig 2.6). The extremely large size of our dataset allows for the detection of very small changes, such that 64% of main CAGE peaks show some statistically significant change. However counting only those CAGE peaks with a log2 fold change of 1.5 or more between the two time points, we find that 26% of our peaks differ between timepoints. Narrow peaks are dramatically more likely to show up-regulation over development, with 52% of narrow and 20% of

broad peaks showing significant log2 fold change greater than 1.5. The overall proportion of genes whose log2fold change is positive is 62.7%. Furthermore, an analysis of GO terms enriched for high fold change between timepoints (see materials and methods) shows that genes with functions in reproduction and development are enriched in these differentially regulated genes (Fig. 2.7). These results again illustrate the nature of the separate gene regulatory systems governing broad and narrow peaks – narrow peaks are highly enriched for genes involved in the development and function of specific tissues, while broad peaks are enriched for genes with more general, less tissue specific functions.

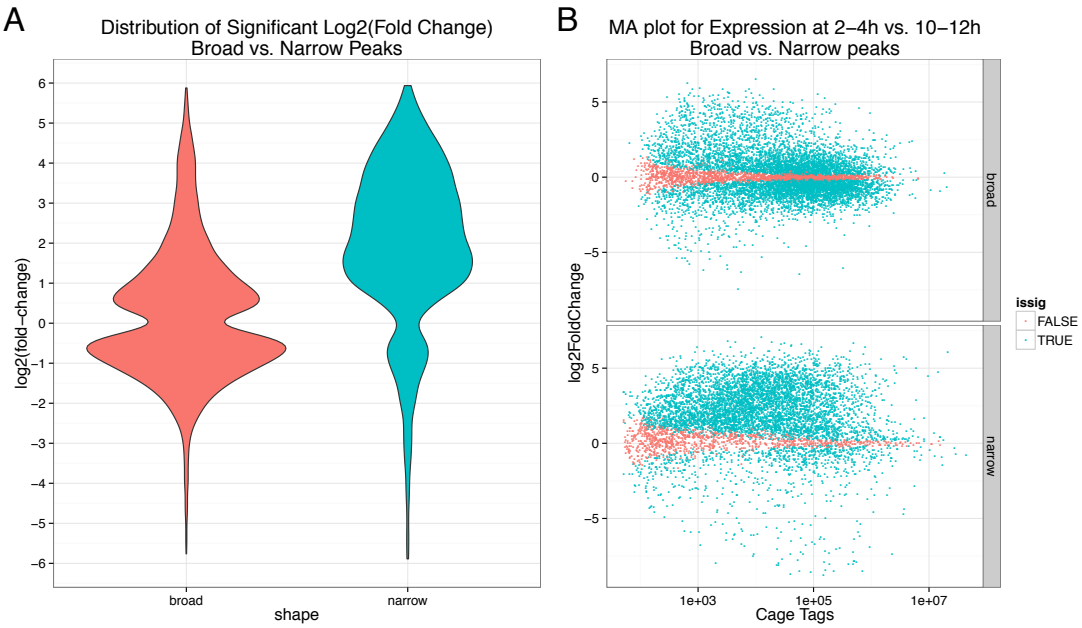


Figure 2.6: Up-regulation of narrow peaks.

A) Distribution of significant ($p < 0.05$) log2 fold changes between 2-4 and 10-12 hours for CAGE peaks in the main set, estimated using DESeq2, for broad (SI < -0.5) and narrow peaks. Note that narrow peaks are overwhelmingly likely to be up-regulated during development, and to a greater extent, than broad peaks. B) Log2 fold changes for all peaks in the main set, versus tag count. Note that the very high depth of our data allows for even small fold changes to be detected as statistically significant.

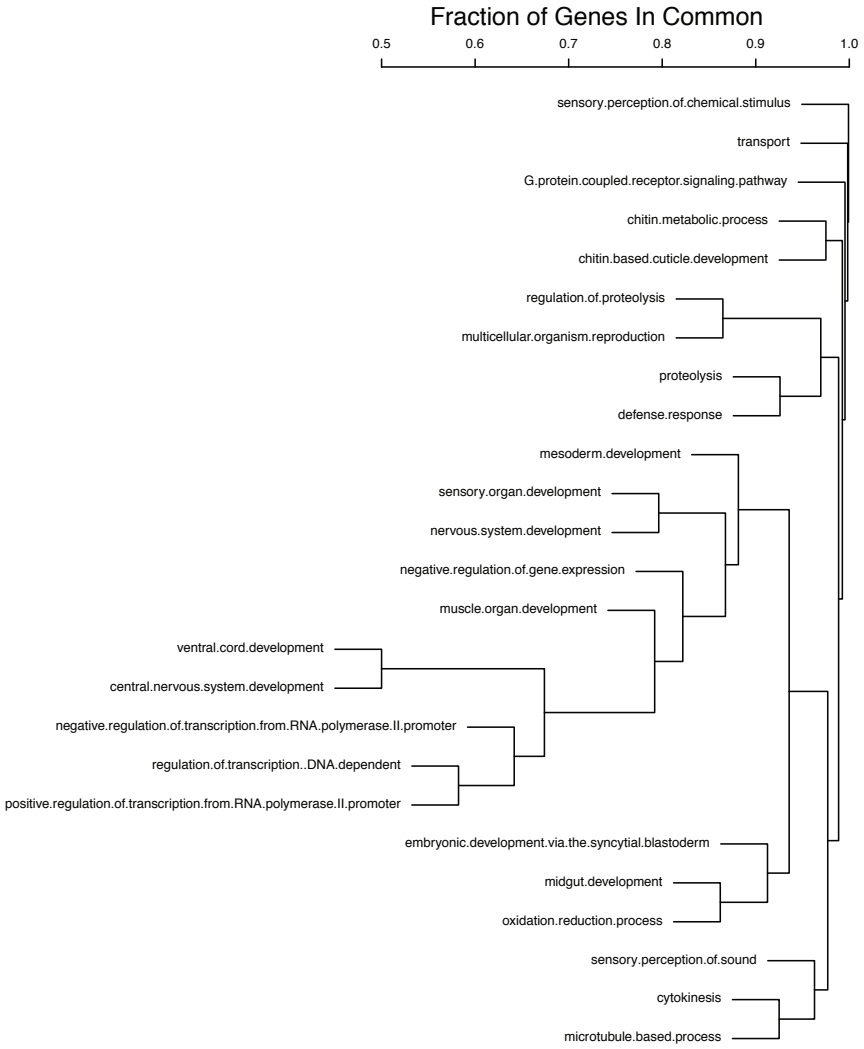
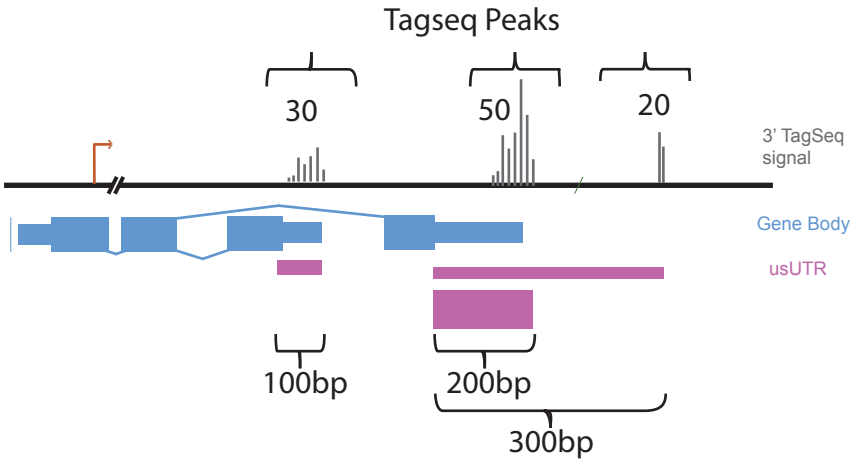


Figure 2.7 GO terms associated with differentially regulated peaks

GOslim terms associated with development are differentially regulated over development. GO biological process terms associated with differentially regulated CAGE peaks. Absolute - log2fold changes were calculated for CAGE peaks using DESeq. A linear model was used to control for gene expression and length, and the residuals from this model were used as the input data for a point by serial correlation analysis. GOslim terms with more than 50 member genes significantly enriched for absolute log2fold-change were plotted. Clustergram reflects the fraction of genes shared by any two categories, with 0 indicating nested categories and 1 reflecting orthogonality.

2.2 3’ UTR biology in the developing embryo

Our study of 3’ Tag-seq QTL used expression data collected in 80 DGRP lines (largely overlapping with the 5’ CAGE lines), collected over three different timepoints, as with the 5’ CAGE data. The data were aligned to personalized genomes and used to derive a set of polyadenylation (pA) sites for QTL calling. Briefly, these peaks were called using reads corresponding to transcript ends, and represent locations where 15 or much such reads were clustered in a single location, with closely spaced regions being merged. The total number of peaks called was 37,025, which more than doubles the previously documented number (Brown *et al* 2014) of pA sites in *Drosophila* . We were interested in using these peaks to examine the biology of 3’ UTR processing. This presents a difficulty in that reads from the 3’ Tag-seq assay are present only at the 3’ end of genes, so that the splicing structure of transcripts associated with a peak cannot be inferred. We therefore elected to define a set of ‘unspliced 3’ UTRs’ (usUTRs) which represent the interval between each pA peak and the presumptive gene of origin’s most likely stop codon (See Fig 2.8), a procedure similar to that used by Sandberg *et al* (2008).



Mean usUTR length = (100bp x 0.3) + (200bp x 0.5) + (300bp x 0.2)
= 190bp

Figure 2.8 Mean unspliced UTR length – definition.

pA sites, which are linked to genes by proximity and strandedness, are linked to the nearest upstream codon, to within a maximum distance of 10kb. The resulting usUTRs are then weighted by their expression values at a given time point, to give an estimate of the mean UTR length for all of the gene's transcripts at that time point.

2.2.1 Distribution of pA sites in *Drosophila*

Having identified usUTRs in *Drosophila* , I wished to examine their distribution and frequency across the genome. In general the variation in the number of pA sites is large, with 10349 (67%) of genes having between one and 68 pA sites linked to them. (Fig 2.9a) Our procedure gave putative usUTRs for most pA sites, with 93% of them occurring less than 10kb downstream of a candidate stop codon, and thus allowing unspliced UTRs to be inferred (Fig 2.9b). Genes had up to 44 usUTRs attributed to them. Most usUTRs were relatively short (See Fig 2.10) with a median length of 399bp and a small number of them positioned large distances from their stop codon (7% were between two and 10kb long). We also observe, in agreement with others (Smibert *et al* 2012), significantly higher UTR lengths in genes expressed in the nervous system (Fig2.10).

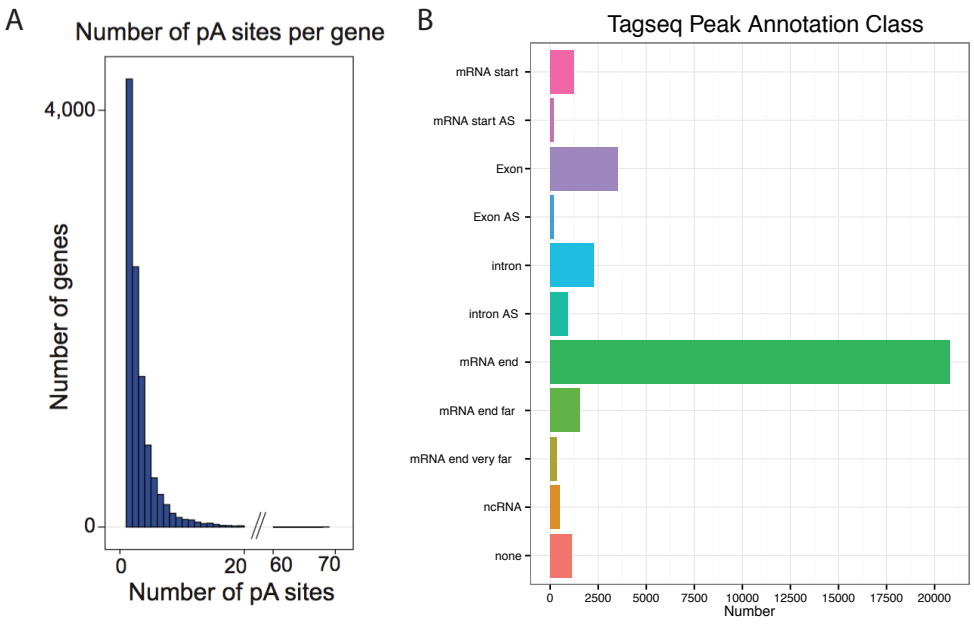


Figure 2.9 Distribution and number of pA sites.

A) Histogram showing the number of pA sites attributed to genes at all stages. B) The various classes into which pA sites were classified. The majority of pA sites fell within 500bp of an annotated transcript end (mRNA-end). Those falling further than 2kb away were labeled mRNA-end-far. pA sites at the start of genes (mRNA start) were antisense to the gene (intron AS, exon AS, mRNA start AS) or which were not attributed to coding genes (ncRNA, none) were not used to construct usUTR.

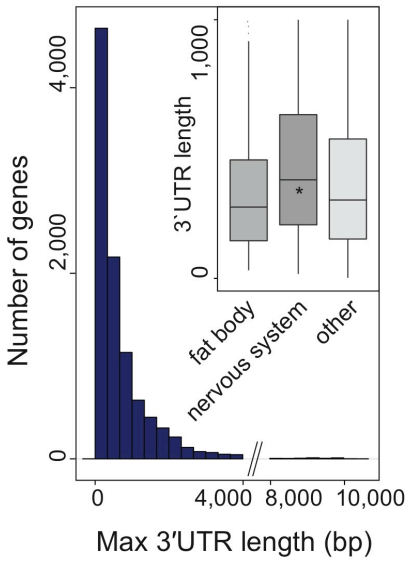


Figure 2.10 3' UTR length and expression pattern

Histogram shows distribution of maximum 3' UTR length per gene over all stages. Inset: boxplots with UTR lengths for genes with expression (see materials and methods) in fat body, nervous system and other tissues. Genes with expression in the nervous system show significantly (Wilcox signed rank test) higher UTR length.

2.2.2 3' UTR length change and expression in *Drosophila*

Since each of our usUTRs was associated with a measure of expression, we were able to examine changes in their relative usage over development. A gene's UTR length cannot be expressed in a single number in cases where more than one UTR exists. Naively, one might suppose that a gene could switch totally between a longer and shorter UTR over development, but in practice, we observed that changes in UTR usage are almost always quantitative, rather than qualitative. We therefore summarized the length of the UTRs in use by a gene at a given time point by weighting each of the usUTR lengths by their relative expression, to derive a 'mean expressed UTR length for each gene. Most genes show relatively little change, (Fig 2.11) with just 6.6% of genes showing a two-fold or greater decrease, and 3.3% showing a two-fold or greater increase. We observed that genes whose expression goes down between 2-4 hours and 10-12 hours, as per a published whole embryo RNA-seq dataset (Brown *et al* 2014), tend to show increases in 'mean expressed UTR

length', while genes whose expression goes up show the opposite trend (Fig 2.12). This indicates that our summary statistic captures the changes in UTR usage known to occur during *Drosophila* development (Smibert *et al* 2012), in which lengthening UTRs are associated with decreasing gene expression, in particular in the developing nervous system, where this change in UTR usage affects the usage of RNA regulatory motifs. We note however that this signal could also be caused by the false attribution of pA sites to genes, since such misattributed peaks would be less likely to show co-regulation with the gene's other pA sites. Reasoning that Gene-pA site attributions at greater than 2kb are the least reliable, we excluded these from the analysis and found that the results remained significant.

2.3 Enhancer transcription in Drosophila

2.3.1 Enhancer transcription in human vs. *Drosophila*

With the advent of high sensitivity expression assays, it has become clear that the distinction between a promoter element and enhancer elements is not as clear as previously thought. Numerous studies have shown that many enhancer elements also have transcription (e.g. Kim *et al* 2010, e.g. Andersson *et al* 2014), however to date, the phenomenon has received only cursory study in *Drosophila* .

While some papers have referenced the presence of transcription at enhancers in *Drosophila* (Kharchenko *et al* 2011, Core *et al* 2012) none, to our knowledge, have addressed the question of whether this transcription is of similar intensity to that in humans.

I began my analysis by examining data from a previous study that had measured transcription at *Drosophila* enhancers (Core *et al* 2012), and compared it to transcription at human enhancers. The authors performed GRO-seq, which measures nascent transcription, on S2 cells, and compared it to an existing dataset collected in human IMR90 cells. They observed that a) transcription at *Drosophila* enhancers was present, and b) that transcription at *Drosophila* enhancers seemed to be more directional than transcription at human enhancers. On visual inspection, I

found that the enhancers used by this paper were very frequently in close proximity to genes, and frequently overlapped a newly annotated TSS, lincRNA, or the region of GRO-Seq signal that often extends past the end of transcribed genes (Fig 2.12).

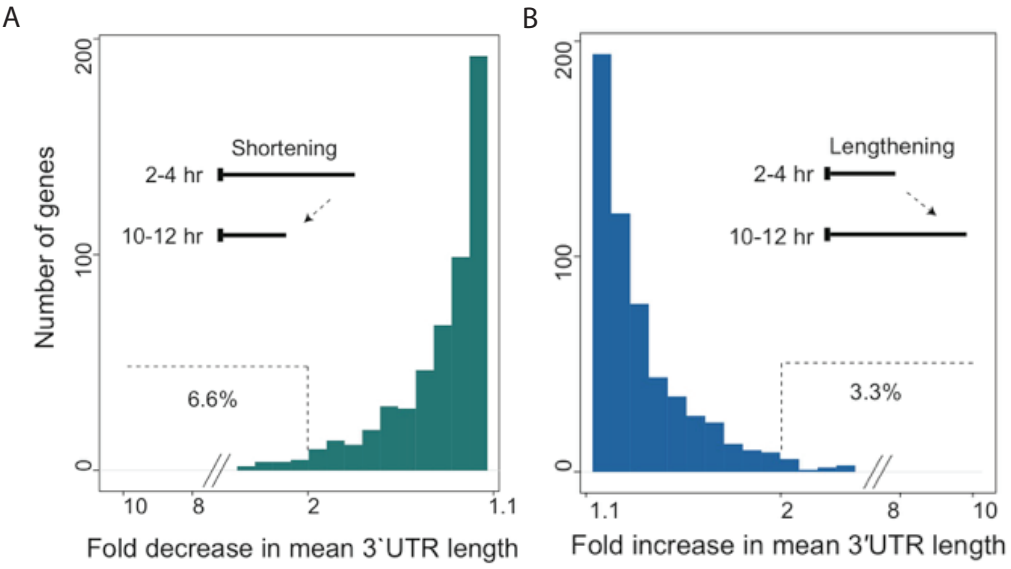


Figure 2.11 3' UTR lengths and length change over development.

Distribution of mean usUTR length changes during development. Mean 3' UTR length is calculated by weighting each usUTR's width by the expression of its pA site at the relevant time point. Histograms show the number of genes either shortening (A) or lengthening (B) the mean length of their transcribed usUTRs.

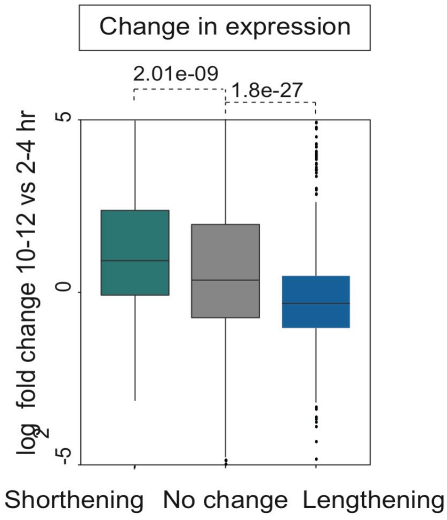


Figure 2.12 usUTR Length changes vs. expression over development.

Boxplots show change in mRNA levels (log2 fold-change of FPKM at 10-12hr vs. 2-4hr within +/-5) for all genes with shortening or lengthening (>10% change) of their mean usUTR length (see Fig 2.11). Lengthening is associated with a statistically significant decrease in expression over time, with the 'no change' group differing significantly from both the shortening and the lengthening group (two-side t-test).

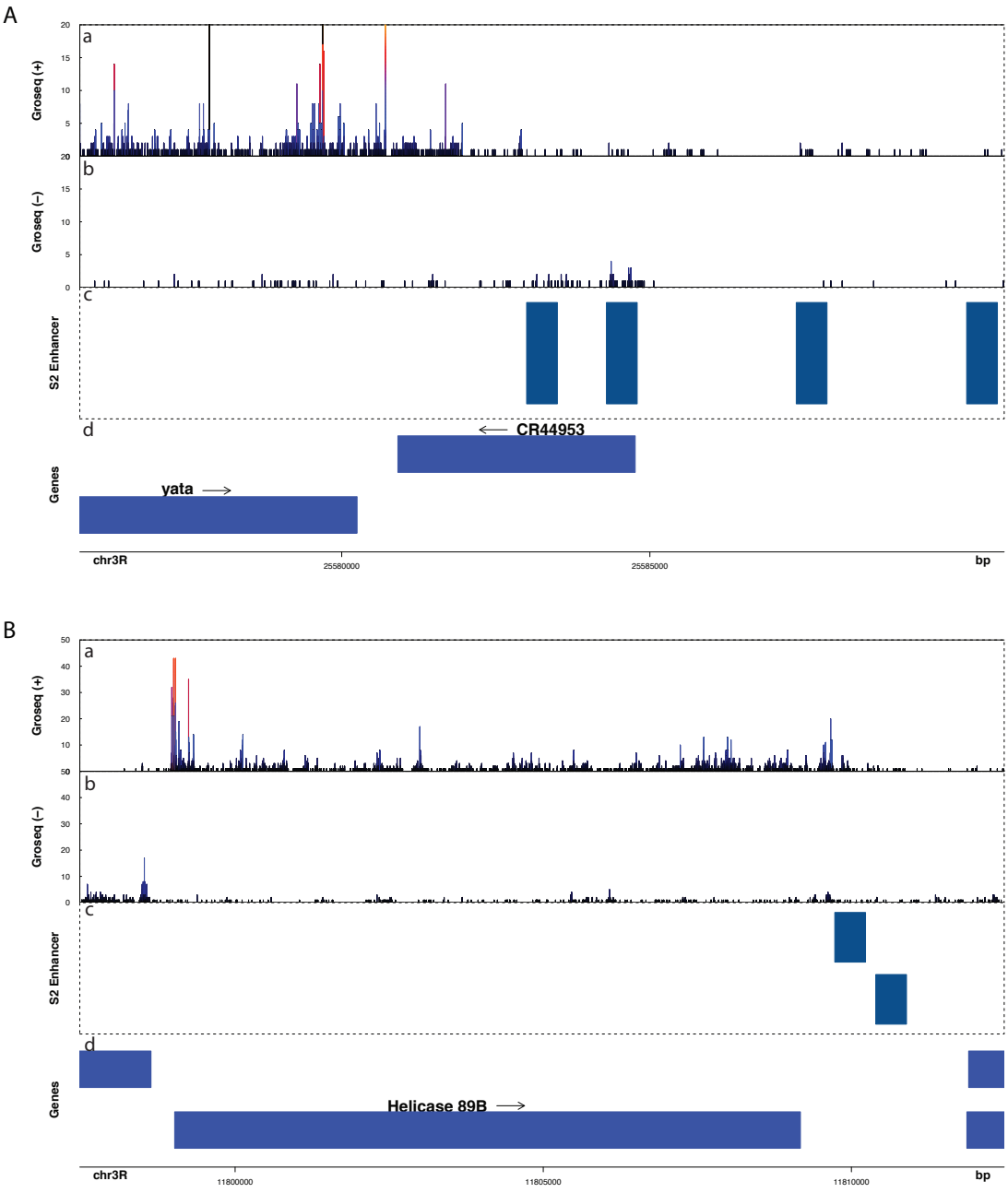


Figure 2.13: Re-annotating *Drosophila* eRNAs

Much of the enhancer transcription identified by Core *et al* is either associated with genes, or with newly annotated eRNA. In order, tracks show GROseq signal for positive strand, negative strand, the locations of Core *et al's* enhancers, and the location of genes. A) Example of an enhancer that overlaps a non-coding RNA. A) The promoter of the ncRNA *CR44953*, located downstream of the gene *yata*, was included in Core *et al's* enhancer set. Shown is the Gro-seq signal for positive and negative strand (a ,b) the enhancers (c) and the gene model (d) B) Example of an enhancer with gene associated transcription. The enhancer immediately 3' of the gene *helicase 89b* overlaps the GROseq signal associated with the terminal end of the gene. Panels (a-d) are as above.

A number of recent papers (Brown *et al* 2014, Batut *et al* 2014,Young *et al* 2013) have significantly increased the number of annotated ncRNAs and transcripts in the *D. melanogaster* database. It appears that many of the enhancers examined by Core *et al* either overlap ncRNAs that were at the time not annotated, or are in close proximity to genes. In order to screen such enhancers from our dataset and thereby examine transcriptional activity in associated with enhancer activity, as opposed to gene activity, we screened out all enhancer which were up to 500bp upstream of an annotated gene, or up to 2kb downstream of one. With these screens applied, we found that *Drosophila* enhancers (in contrast to *Drosophila* promoters) were no more directional than human enhancers (Fig 2.13) indicating that this finding by Core *et al* was due to the inclusion of gene-associated transcription within their enhancer set.

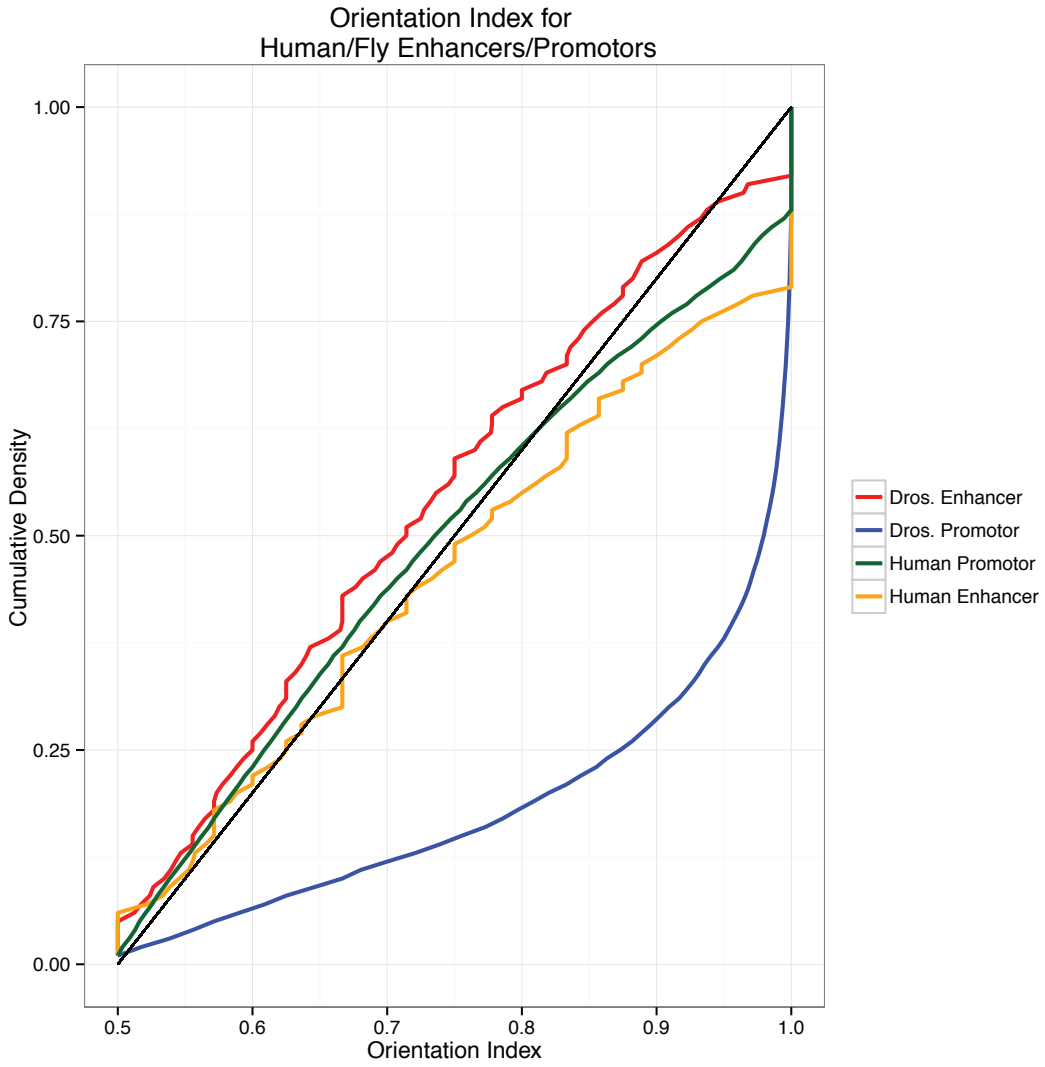


Figure 2.14: Orientation index for human And *Drosophila* regulatory elements.

The directionality index used by Core *et al* (2012) (the total reads on the dominant strand as a fraction of the total reads, x-axis) was used to examine the strand bias seen in regulatory elements. The diagonal represents the distribution of reads expected for regions with a directionality index distributed uniformly (as would be expected if the two strands at a given region vary independently in their activity). The blue line indicates that *Drosophila* Promoters are notably more directional than human promoters, or enhancers in other species. In contrast to Core et al, we find that *Drosophila* eRNAs show a similar distribution of directionality to human eRNA (and human promoters).

We also wished to ask about the magnitude of eRNA expression in human vs. *Drosophila* . The comparison is a relatively difficult one to make because of the lack of any 'standard candle' that could have allowed us to link GROseq levels to cellular RNA concentrations. There is furthermore no reason to expect the relevant scaling factor to be proportional to the number of genes, number of enhancers, or genome size of the two organisms. I therefore made the approximation that considering all expressed genes; the median expression should be similar between S2 cells and IMR90 cells. I then resampled the IMR90 GROseq data so that the median levels for expressed genes were equal to those in the S2 data (Fig 2.15a). This allowed for a comparison of the GRO-seq counts in the two sets of enhancers. Fig 2.15b shows the result – transcription from enhancers in *Drosophila* S2 cells is of comparable magnitude to that in human IMR90 cells.

These results suggest that enhancer transcription is present in *Drosophila* cells in roughly similar quantities to human cells, and that it is also qualitatively similar to – rather than being more directional than, human eRNA.

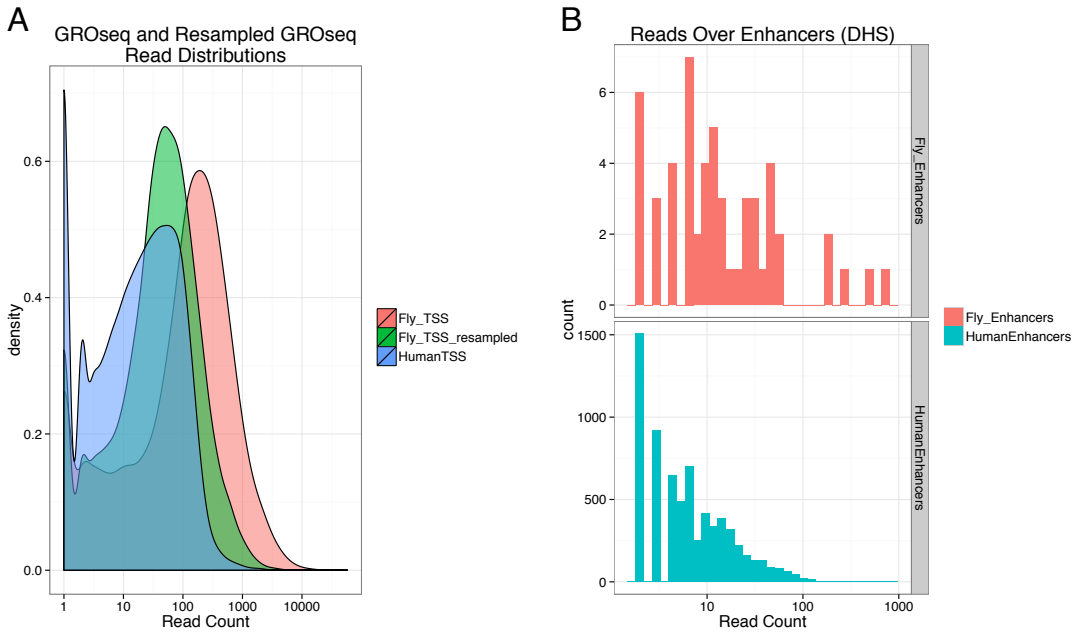


Figure 2.15 Comparing magnitude of eRNA expression in *Drosophila* S2 cells and human IMR90 cells.

A) Density plot showing read counts over 17,256 annotated TSS for Fly (Flybase dm3) and 57,736 for Human (Refseq h19). GRO-seq reads within the 1kb window surrounding TSS were counted (x-axis), taking the best scoring TSS for each annotated gene. *Drosophila* GROseq data from Core *et al* 2012 (red) was resampled (green) until the median count over annotated TSS was equal to that in the human GRO seq data (blue), to derive a scaling factor for comparison of enhancers. B) Comparison of GROseq counts for the 77 (filtered) *Drosophila* enhancers and 14,408 human enhancers. Note that stringent filtering on the *Drosophila* enhancers was required to exclude gene associated transcription - only small numbers of *Drosophila* enhancers remain after filtering those proximal to genes or newly annotated ncRNA. Reads within the 1kb zone around enhancers (as detected by relevant chromatin marks) were counted using the human and fly data, resampling using the scaling factor derived in the previous step.

Having established the presence of eRNA in *Drosophila* , we wished to examine its association with enhancer activity. To do this, we first took advantage of published (Arnold *et al* 2013) STARR-seq dataset and DHS datasets. STARR-seq is a high throughput means of assaying enhancer activity in cultured cells that measures the ability of a given segment of DNA to drive expression of a reporter assay. While the assay is in principle applicable to humans as well, only small (~1MB) segments of the genome have so far been tested, since a much greater amount of sequencing would be necessary. The STARR-seq dataset thus represent a tool available only for *Drosophila* , and is particularly relevant to the question of eRNAs and their functional relevance, since it directly measures regulatory activity rather than some biochemical proxy such as DNase sensitivity or the presence of chromatin marks, albeit in the altered context of an expression plasmid. To do so, we took advantage of the many different expression assays available for S2 cells. A variety of means for detecting eRNAs have been used (reviewed in Murakawa *et al* 2016). Assays like GROseq that measure only nascent transcription have the advantage of being able to detect unstable transcripts, while assays which assay only TSS like CAGE are useful in that they allow easier identification of eRNAs, whose location and length are difficult to determine. PROcap therefore combines the advantages of both assays. Meanwhile the assay developed by Kwak *et al* 2013, which sequences short, capped, nuclear RNAs, should in principle yield a very similar signal to that from PROcap (since enriching for nuclear RNA will also enrich for nascent RNAs) without the need for difficult and costly nuclear run on assays. Applying the same screens for gene overlap we applied to Core et al's enhancers, we examined signal over DHS in S2 cells, reasoning that this should allow reasonable comparisons with the enhancer DHS based sets used by e.g. Andersson *et al* 2014. Comparing available datasets for each of these four expression assays, (Fig 2.16,2.17),

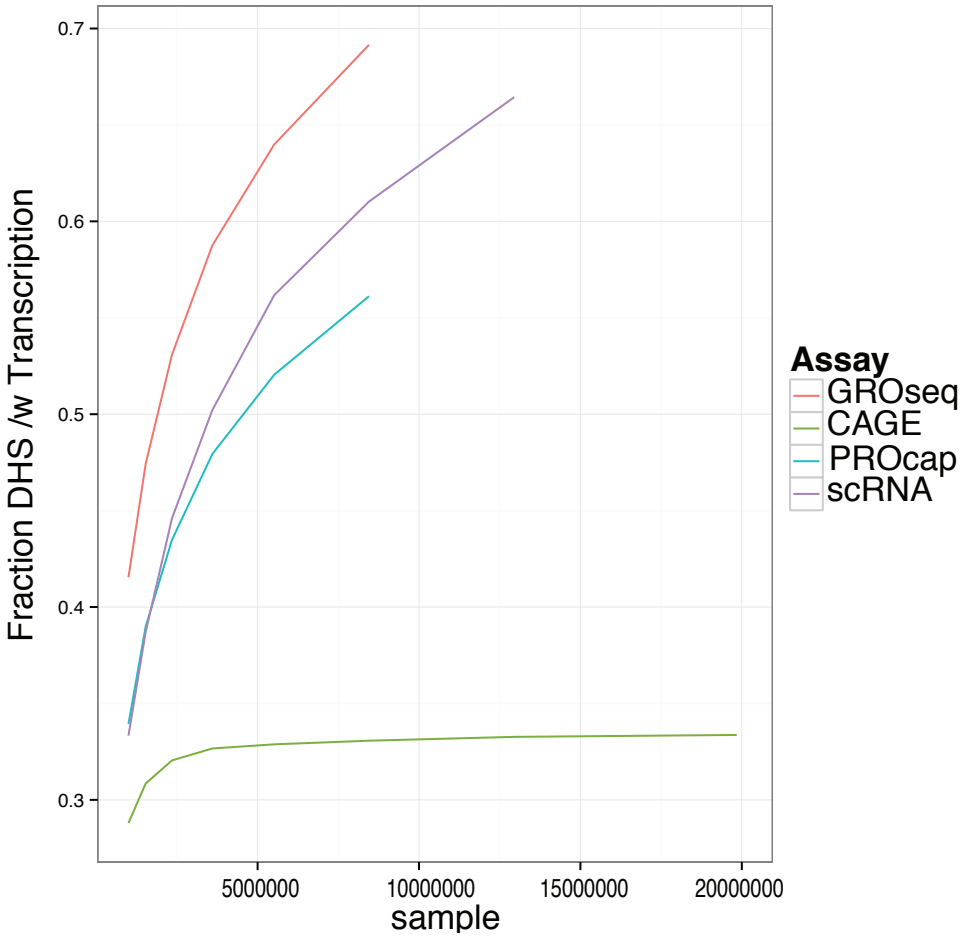


Figure 2.16: Saturation analysis for various expression datasets in S2 cells.

Fraction of S2 DHS with expression (y-axis, defined as one or more reads for the relevant assay) vs. the number of reads resampled from the data (x-axis) for various expression assays. The CAGE (green) data, which shows expression for few DHS, but has 46M reads, has reached saturation. The other datasets show far more DHS with expression but are clearly of insufficient depth to detect all eRNA.

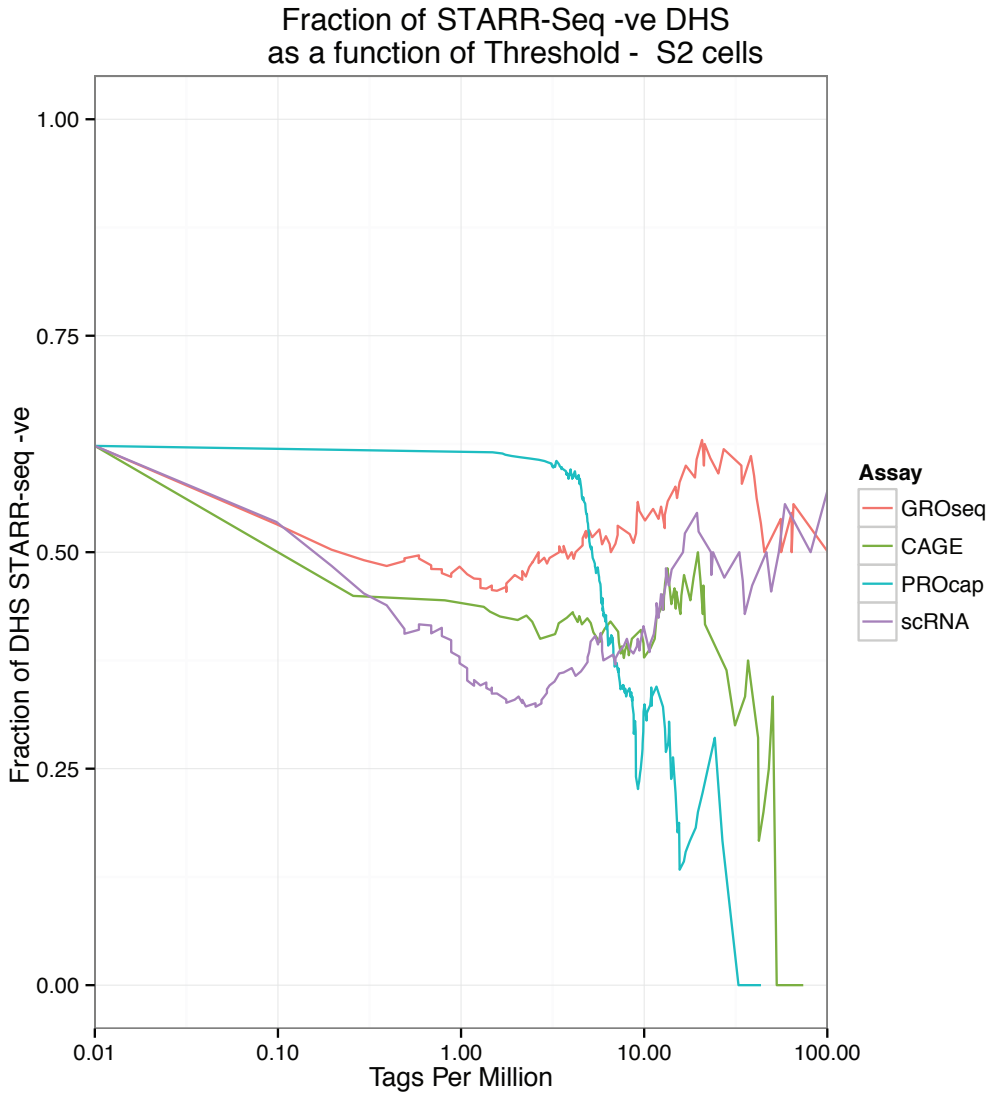


Figure 2.17 False positive rate vs threshold for expression assays – S2 cells.

Fraction of S2 DHS with without STARR-seq activity (y-axis) as a function of the expression threshold used in Tags per Million (TPM) (x-axis), for various expression assays. The proportion of DHS which test positive for enhancer activity initially rises, but then begins to fall again (after ~ 1 TPM) for GROseq (red) and scRNA (purple), which is likely a result of the fact that higher tag counts can also reflect that the DHS is within a transcribed region. This pattern is not evident for CAGE (green) or PROcap (blue), which are present only at TSS.

I found that many were of insufficient depth to detect all transcription at DHS, the exception being CAGE data, which while of very high depth, nevertheless showed expression at few DHS, likely reflecting the instability of eRNAs and the fact that CAGE measures steady state RNA levels. GROseq, scRNA and PROcap all showed a higher proportion of expressed DHS, but resampling shows that they had not yet saturated the eRNA signal.

I then compared DHS in S2 cells with and without STARR-seq activity, using the STARRseq +ve DHS as a set of validated enhancers, and STARR-seq –ve DHS as a set of negatives. Assessing the false positive rate of various expression thresholds (Fig 2.17) shows that the false positive rate (fraction of STARRseq –ve DHS) reaches a minimum at 1-10 TPM for both GROseq and scRNA, before climbing again at higher thresholds. This may reflect the fact that very high levels of short capped RNA or GROseq are often reflective of DHS within transcribed regions, since PROcap and CAGE, which do not exhibit the pattern as clearly, give signals focused exclusively on TSS.

We find also find that, as in humans, there is an association between enhancer activity and transcription, with STARR-seq active DHS showing significantly more GRO-Seq or sncRNA-Seq reads than inactive DHS. For instance, Andersson *et al* found that at a lenient threshold of 0.5 Tags per Million, the validation rate of DHS in enhancer assays climbed from 27% to 57%. Our results in S2 cells are similar, with for instance an increase from 20% to 60% (OR = 5.8, p=3.59e-11 fisher's exact test) in the validation rate, applying the same threshold to scRNA, and more modest increases enrichments for CAGE, GROseq and PROcap (Fig 2.19). Notably, however many DHS show no STARR-seq activity but considerable transcriptional activity, and vice versa (Fig 2.20).

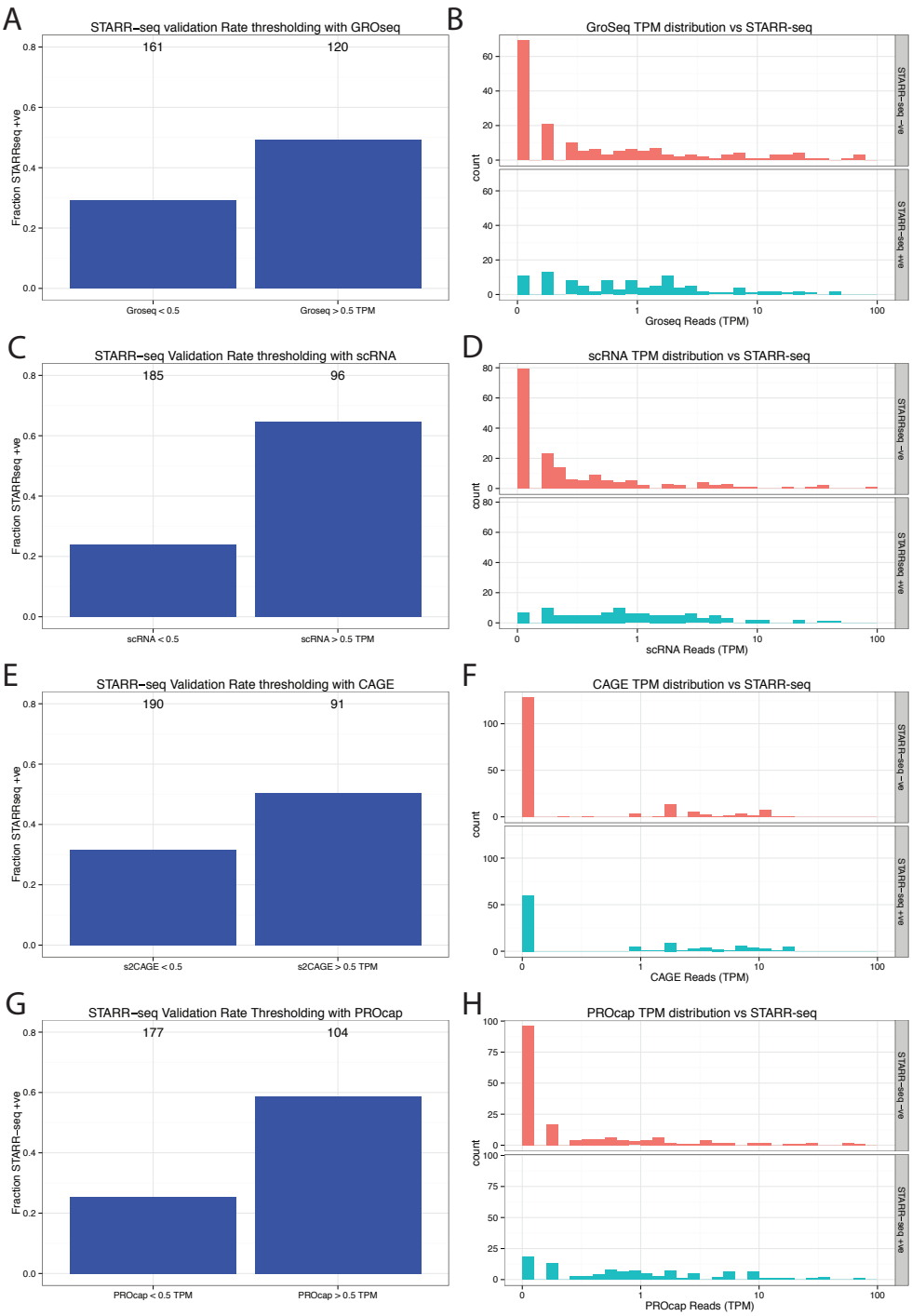


Figure 2.18 – distribution of expression signal over extragenic DHS in S2 cells.

A,C,E,G) The fraction of intergenic (>1kb from a gene) S2-cell DHS overlapping STARR-seq peaks for peaks above and below a GROseq (A), scRNA (B), CAGE (C), or PROcap (D) threshold of 0.5 TPM (Tags Per Million). B, D, F, H) The distribution of overlapping read counts, in TPM, for GROseq (A) scRNA (B), CAGE (C), and PROcap (D). Note that transcription has some predictive power with respect to enhancer activity as measured by STARR-seq.

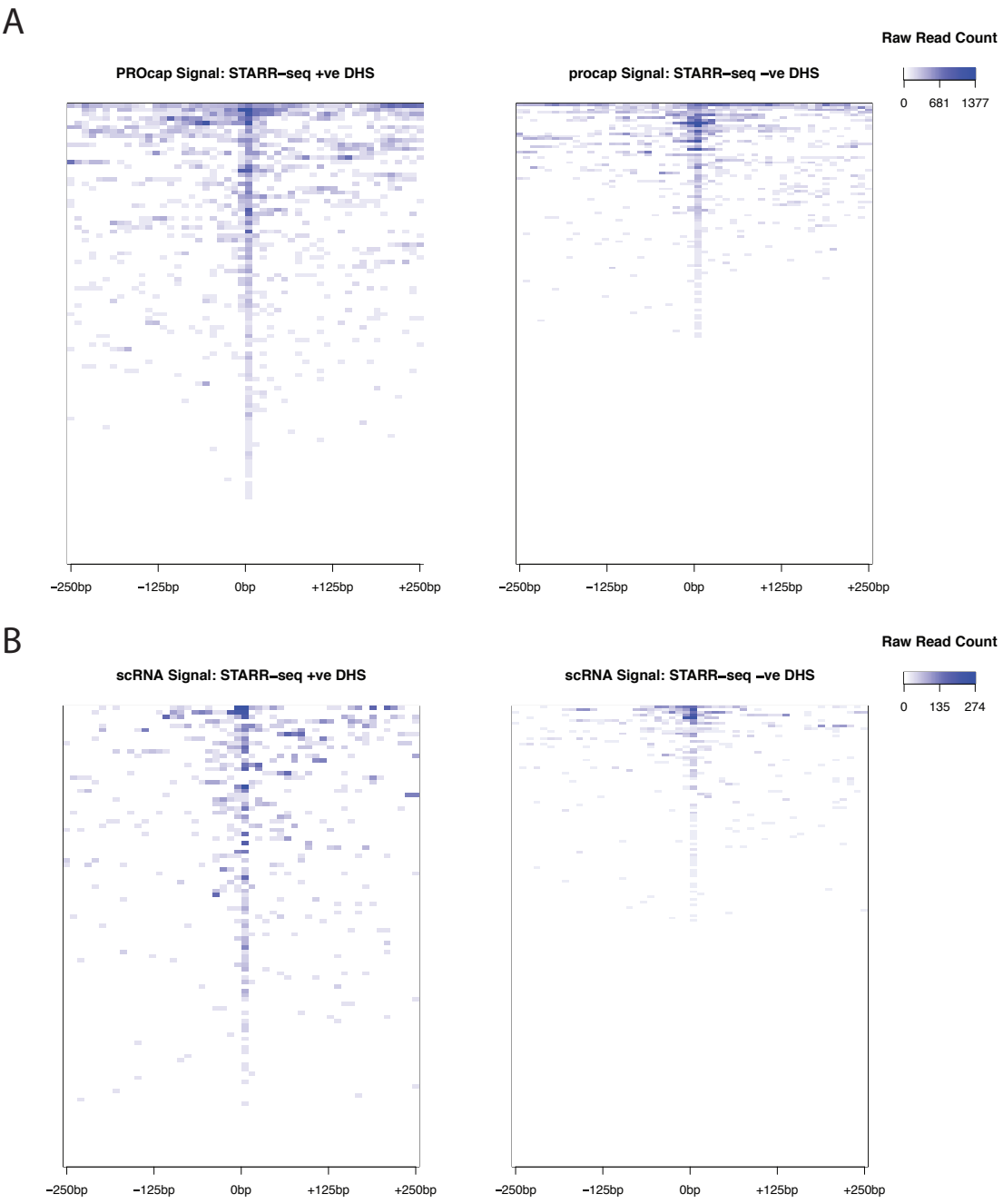


Figure 2.19 Signal over DHS in S2 cells

Heatmap showing distribution of reads over intergenic (>1kb from a gene) S2-cell DHS using PROcap (A) and scRNA (B). Shown is the 500 bp window centered on the point of highest expression for each DHS. STARR-seq +ve DHS (left) overlap at least one STARR-seq peak, indicating enhancer activity, while STARR-seq negative DHS do not. Not that both positive and negative sets show some transcriptional activity.

2.3.2 Enhancer transcription in the *Drosophila* embryo

Drosophila remains a key model organism in the study of gene regulation in large part because of the ease of doing developmental biology on its fast developing embryos. We wished to examine the presence of enhancer transcription in whole embryo datasets, to see if the presence of eRNA and its association with activity remained visible. To do so, we turned to a large database of experimentally verified enhancers collected from the literature (see materials and methods). Since this database consists of relatively large regions, we overlapped our enhancers with DHS taken from a study of DNase sensitivity in whole embryo (Thomas *et al* 2011) (see materials and methods), assuming that the DHS should generally represent the active sequence within each tested region. This dataset affords us the ability to distinguish between a) active enhancers - enhancers which are active at a given time point b) inactive enhancers - enhancers which are inactive at a given time point but are active at some other time point, and c) 'negative' regions which have been tested for enhancer activity and were found to be inactive at every time point. The latter represent an important control set that have thus far been missing from studies of eRNA. To examine transcription of enhancers in the developing embryo, we focused on a single time point – 6-8 hours, and made use of both the 5' CAGE data used to call CAGE peaks, and a PROcap dataset collected at the same time point in whole embryo. Saturation analysis showed that, again, only the CAGE data was fully saturated, due to the small number of reads in the other datasets, and that unlike in S2 cells, at saturation 100% of DHS showed at least one CAGE tag. In agreement with our findings in S2 cells, we find that there is a moderate association between enhancer activity and enhancer transcription (Fig 2.21).

Using DHS overlapping our enhancers as true positives, and DHS overlapping inactive or negative regions as negatives, we plotted false positive rate as a function of the threshold used to define eRNA (fig 2.21). This showed that a slight decrease in FPR occurs for both CAGE and PROcap, with CAGE showing a somewhat greater decrease. Applying a threshold on CAGE and PROcap at the same level used in S2 cells (0.5 TPM) gave no significant difference in numbers of DHS overlapping enhancers, however a lower threshold of 0.32, chosen by the inflection point of the

FPR plot (Fig 2.21), gave significant, albeit lower, association between expression and enhancer activity (Fig 2.22).

As in S2 cells, plotting signal over DHS shows that that even some regions within the 'negative' set nonetheless show transcription (Fig 2.23). Notably, inactive and negative enhancers show similar expression levels. Previous reports (Wu *et al* 2014) have indicated that eRNAs undergo a baseline level of expression before being unregulated in concert with enhancer activity. Our data indicates that this 'baseline' level may simply be reflective of the background levels of eRNA expression in open chromatin, since inactive enhancers and negative enhancers are transcribed at similar levels.

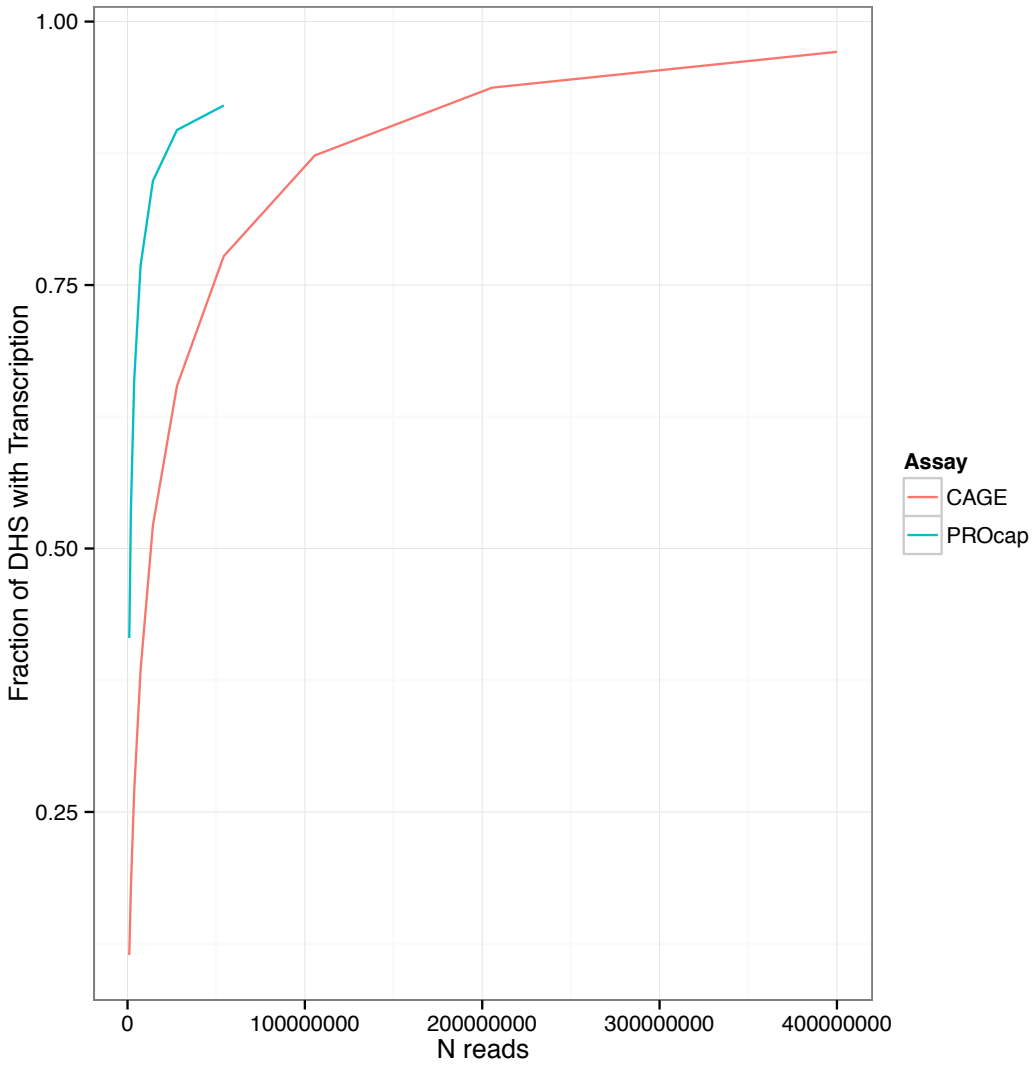


Figure 2.20 False positive rate vs threshold for expression assays – whole embryo

Fraction of whole-embryo DHS with expression (y-axis) vs. the number of reads resampled form the data. The CAGE data shows expression in almost all DHS given sufficient sampling depth, while the PROcap data appears to be undersaturated.

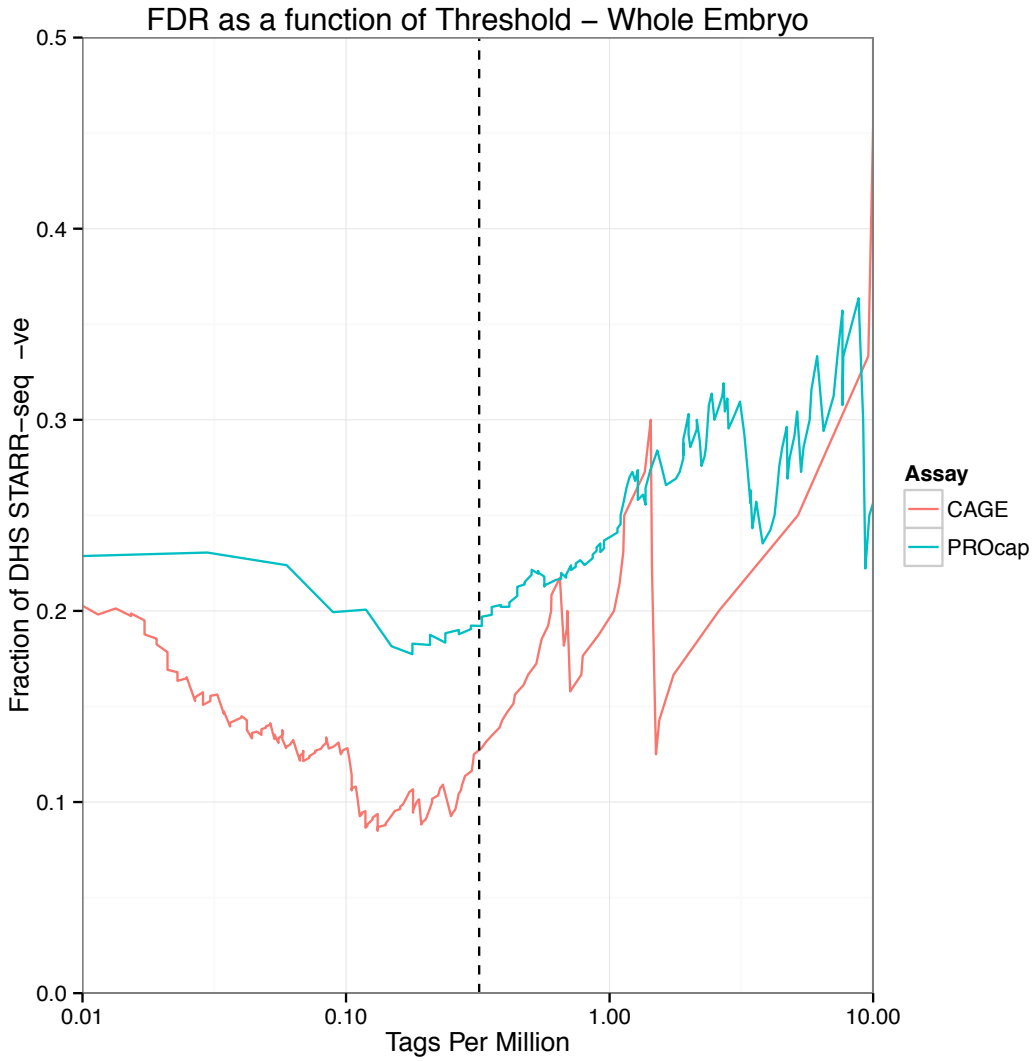


Figure 2.21 False Positive Rate vs threshold for expression assays.

Fraction of whole-embryo DHS without expression (y axis) vs the number of reads resampled form the data. Here 'the false positive rate' shows far less variation than in S2 cells, with most DHS overlapping a transgenically active enhancer. The much greater depth of the CAGE data likely explain its better performance.

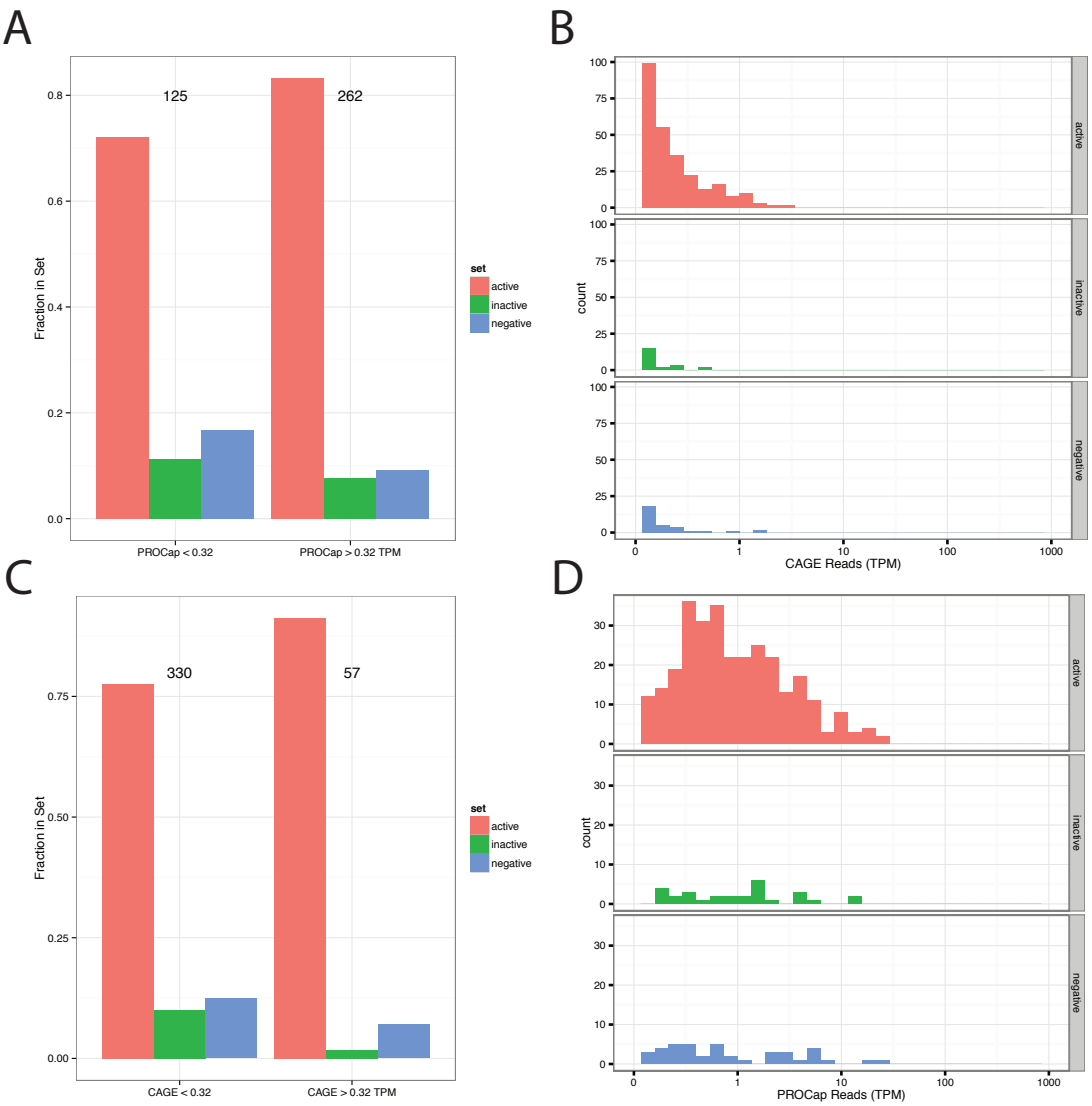


Figure 2.22: Distribution of Expression signal over extragenic DHS in whole embryo cells.

Shown is increase in validation rate when thresholding at 0.34 TPM (Tags Per Million) (A, C) for active,inactive, and negative enhancers. DHS are classed as either active (red), inactive(green) or negative (blue). FPR and odds ratios were calculated by treating inactive and negative enhancers as equivalent. Also shown is the distibution of read counts over each class of enhancer (B, D).

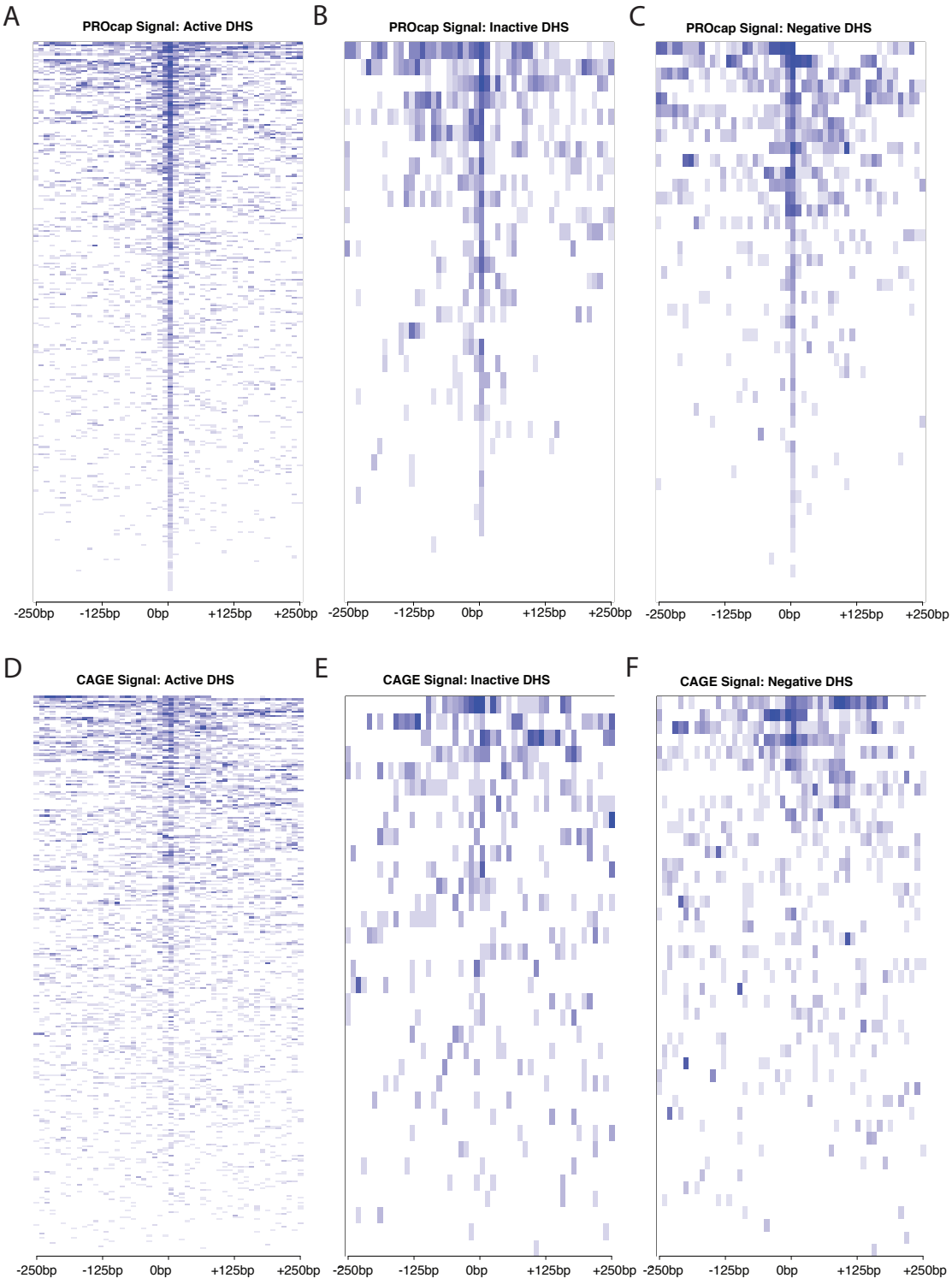


Figure 2.23 Heatmap showing distribution of transcription using PROcap and CAGE for DHS in whole embryo.

Heatmaps showing signal distribution for PROcap (A-C) and CAGE (E-F) over various classes of DHS in whole embryos. Shown is the 500 bp window centered on the point of highest expression for each DHS identified at 6-8 hours. DHS on the left (A, D) overlap a region annotated as active at 68h via transgenic reporter assay. DHS in the

middle (B, E) are active at other timepoints, but not 68h. DHS on the right (C, F) have been tested and found not to have any activity at any timepoint.

2.3.3 Testing enhancer transcription *in vivo*

We wished to determine whether the regions we identified as having enhancer transcription *in vivo* would also drive detectable tissue specific activity of a reporter gene *in vivo* . In order to confirm that our enhancer RNAs possessed specific transcriptional activity *in vivo* , we decided to carry out transgenic reporter assays. In order to select candidates for validation, I selected intergenic regions bound by at least three mesodermal transcription factors (Zinzen *et al* 2009), and displaying the histone modification H3K27ac, and RNAPII binding, at 6-8 hours (Fig. 2.24). We reasoned that enhancers with highly stable eRNA (high CAGE/PROcap) might be more likely to show detectable transcription than those with less stable eRNA. We chose four enhancers with a high ratio of CAGE to PROcap, and four enhancers with a low ratio of CAGE to PROcap. As expected, most (7/8) enhancers displayed tissue specific enhancer activity *in vivo* . In 4/8 cases (Fig 2.25, Table 2.1), three from the high CAGE/PROcap group and one from the low CAGE/PROcap group, our enhancers were able to drive tissue specific transcriptional activity when used in place of a promoter. Since a low ratio of CAGE to PROcap should indicate an unstable transcript, it is possible that some of these enhancers also produced unstable reporter transcripts, and simply drove expression at a steady state level too low to detect. The expression pattern shown by the enhancer-only construct, where present, showed a similar, but weaker, pattern to the enhancer-promoter construct. These results demonstrate, for the first time to our knowledge, that enhancers can act as promoters in transgenic reporter assays. Given that only 3/8 of the enhancers with enhancer transcription showed visible expression, it is possible that enhancer transcription is context dependent, and that their insertion into a new genomic context disrupts their normal transcriptional activity.

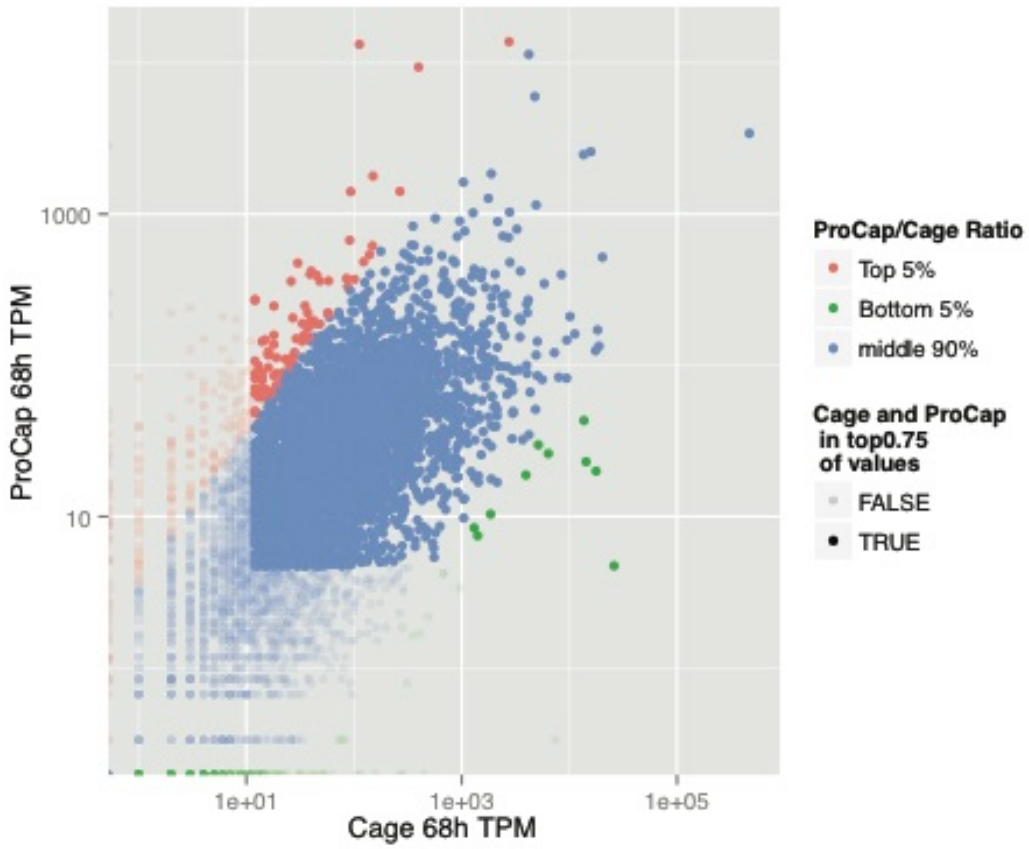


Figure 2.24 Selecting enhancers for validation

Distribution of CAGE (x-axis) and PROcap (y-axis) in TPM (Tags per Million) over the 8008 ChIP identified enhancers from Zinzen *et al* 2009. "Unstable" enhancers were chosen from the bottom 5% of the PROcap/CAGE distribution (green dots), while stable enhancers were chosen from the top (red dots). Enhancers in the bottom 25% of expression for either assay were excluded.

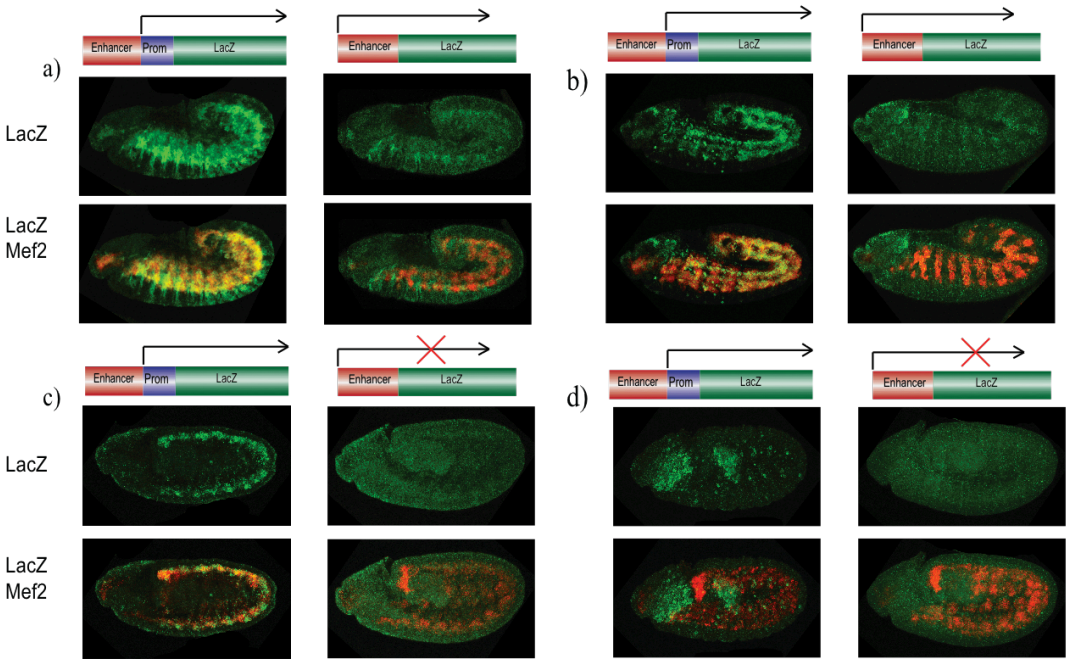


Figure 2.25 In vivo expression validation of *Drosophila* eRNA

Double *in situ* RNA hybridization with probes against reporter gene (LacZ) transcript and mesodermal marker gene (*Mef2*) in different classes of enhancer constructs a), c) – enhancers characterized by high 5' CAGE and PROcap signal b), d) – enhancers that exhibit no detectable 5' CAGE signal and low PROcap. In cases a) and c) enhancer element cloned upstream of the reporter gene is driving expression in the absence of a minimal promoter. In cases c) and d), no visible expression is observed in enhancer only transgenic flies, indicating a lack of detectable 'eRNA' activity in the enhancer.

	Construct	Pol II	5'CAGE	PROCAP	E+P	E only
"Stable"	6v1 (CRM)	+	+	+	+	+
	4d (CAD, BN005lf)	+	+	+	+	+
	3c (CAD, BN005lf)	+	+	+	+	+/-
	8a (CAD, BN31)	+	+	+	+	-
"Unstable"	1b (CAD, rpr4s5)	+	-	+	+	+
	7h (CRM)	+	-	+	+	-
	11t (CAD, mir1_mir1)	+	-	+	+	-
	12c (CRM)	+	-	+	-	-

Table 2.1 In vivo expression validation of *Drosophila* eRNA

Shown are the results of *in vivo* expression assays for all eight tested enhancers. "Stable" enhancers are characterized by high levels of CAGE and PROcap, "Unstable" enhancers have PROcap, but relatively little CAGE signal. Correspondence to Fig 2.25 is as follows: 4d - a, 1b – b, 11t – c, 7h – d.

2.3.4 Discussion

My analysis of CAGE data and 3' Tag-seq facilitates further analyses of motif content and genetic variation, but also complements an increasing number of papers demonstrating that the transcriptome is more complex than has previously been appreciated (Brown *et al* 2014, Young *et al* 2013). A common theme that emerges, and one my work confirms, is that transcriptional complexity is strongly associated with a distinct class of genes, which tend to show spatially restricted expression patterns (Hoskins *et al* 2011, Carninci *et al* 2006, Rach *et al* 2009), narrow promoter shape, and more complex 3' UTRs and pA site usage (Simbert *et al* 2012).

Particularly interesting are the small subset of 'ultra complex' genes that account for a large fraction of the distinct transcripts within metazoan genomes (Brown *et al* 2014). The biology of 3' UTR usage is emerging as an important component of metazoan gene regulation (Simbert *et al* 2012, Hilgers *et al* 2012), and the complex 3' UTRs in these genes may reflect the need for their many isoforms to be independently regulated. My work supports previous indications that changes in the length of 3' UTRs are a mechanism of gene regulation, and that longer 3' UTRs are associated with the nervous system.

At present, almost every paper making use of 5' CAGE uses a different method to define 'peaks'. The reason that no well accepted method has emerged is likely the intrinsic difficulty of the task – unlike say, ChIP data, for which data can be separated into signal and a relatively well defined noise proportional to input chromatin, there is as of yet no well accepted model for the 'noise' which affects CAGE. It is not clear in fact if there is any significant technical noise component within CAGE data – it may be that every location showing a CAGE tag represents a site that is transcribed by PolII at some point. The strengths of individual CAGE tag sites show a power law distribution, such that even in ultra-deep datasets such as ours, the number of weak sites in the genome is nowhere near saturated. More likely, all CAGE datasets contain some small component of contamination from genomic DNA as well. The question of what constitutes a 'significant' CAGE tag peak is therefore a difficult one. Models which attempt to account for noise within transcribed regions, such as my own, will benefit from newer, deeper RNAseq datasets, as well as more sophisticated models of where re-capping and other processes tend to occur (such as ones differentiating between noise levels within exons and introns, and around splice sites). Assessing the significance of extragenic transcription is more difficult, but is of close relevance to studies of enhancer transcription. The continuous distribution of transcription over enhancers presents a serious difficulty to studies of Erna – even if all enhancers show transcription at sufficient sequencing depth, there is presumably some level below which transcriptional activity is unlikely to be significant. Estimating this level will require not only that highly sensitive expression assays be linked to physical measures of molecular concentration, but also that the variance in expression of eRNAs between individual cells be measured. An eRNA present in 10 copies in all cells is more likely to play an important role in enhancer function, than an eRNA present in 1000 copies in 1/100 cells.

I have demonstrated that the phenomenon of enhancer transcription exists in *D. melanogaster* to a similar extent as in mammalian systems. I have also demonstrated that it is associated with enhancer activity, and to an extent that resembles its association in humans. An important caveat emerging from my work is that the dense nature of the *D. melanogaster* genome makes it more challenging to

separate genic from non-genic transcription. The weaker association seen in whole embryo data is likely a result of the tissue specific expression of eRNA. With only a small fraction of cells transcribing a given enhancer, the signal to noise ratio of CAGE or PROcap in whole embryo data is likely less than that in cell culture data. Another, complementary explanation is that DHS data provides more explanatory power in the context of whole embryo regulatory activity than in the context of activity in a single cell, leaving less for transcriptional activity. The non-transcribed DHS in S2 cells could represent regulatory elements active in other contexts, and with more regulatory contexts represented in whole embryo data, DNase sensitivity would more closely represent activity.

Studies making use of tissue specific expression data should be able to go some way towards resolving this question. Studies have shown (Zhu *et al* 2013) that eRNA, when incorporated into models including chromatin data, can improve predictions of enhancer activity. Studies of large numbers of cell lines (Yao *et al* 2015) have also shown that eRNA shows tissue specific activity that co-varies with their target genes. The use of tissue specific assays to collect data on eRNA from whole organisms therefore offers a promising means of assessing tissue specific regulatory activity.

3 The Sequence Determinates of Transcriptional Regulation in the Developing *D. melanogaster* Embryo

Any attempt to analyze the mechanisms underlying eQTL in *Drosophila* will require an accurate means of identifying functional sequences. During my thesis work I made use of existing PWM data, but also used the TSS, 3'UTRs and ChIP peaks I had identified and collated to identify statistically enriched sequences, and carried out quality control on these discovered and known sequences, to yield a high quality set of TSS, pA site, and ChIP-peak associated motifs. Because they are associated to CAGE, 3' Tag-seq, and ChIP datasets by a uniform set of accession IDs and codes, these motifs represent a single unified body of regulatory information to be employed in later analysis. In this chapter, by improving on an extending our knowledge of which sequence motifs are associated with transcriptional regulation in *Drosophila* , I lay the foundation for the motif based analysis of eQTL which is to follow in the subsequent chapter, as well as making substantial contribution to our knowledge of DNA sequence motifs in *Drosophila* .

3.1 Collating existing ChIP and PWM data in *D. melanogaster*

The genomics literature for *D. melanogaster* contains a very large amount of ChIP data. It also contains a very large number of PWMs for various factors. Previous attempts to collate this data have been incomplete (Negre *et al* 2011,Boyle *et al* 2014) and in order to facilitate our analysis of QTL mechanism and enhancer function, we elected to compile an up to date database of all ChIP peaks in *D. melanogaster* embryonic development. This dataset represents an important resource for *Drosophila* genomics in general, and for my thesis in particular. It comprises 424 total datasets for 139 distinct factors at 65 different time points. We ranked our datasets by their source and reliability, taking into account factors such as lab of origin and the technology used to generate the peaks (with e.g. ChIP seq data being prioritized over Chip-chip data). We then took all time points/cell lines for

the highest-ranking group, to form a 'High Quality' category. We also included time points that were not included in the higher quality set, but were in a lower category, under 'Medium Quality'.

In addition to the ChIP data, I also collected a large database of 1025 Position Weight Matrices. These matrices were gathered from diverse sources including the Flyfactor and Jaspar databases, and smaller collections from the Berkeley *Drosophila* Transcription Factor Network Project (BDTNP) and related sources, and the modEncode project (see materials and methods for table of sources).

Each of these position weight matrices was linked to a unique, and up to date Flybase gene accession number, where such a link could be made. Crucially, since both chip and PWM datasets have been given unique and matching Flybase Gene IDs, the two datasets can be easily intersected, something that is often difficult using existing databases, due to the diversity of gene identifiers in use, changes in identifiers over time etc.

Source	Peaks	Factors	Timepoints	Datasets	With PWMs
Furlong Lab	57868	13	7	28	13
modEncode Boyle et al	136619	51	22	83	38
Berkeley Kaplan et al	29391	6	1	6	6
Berkeley BDTNP	126360	25	4	27	23
modEncode - Negre et al	1157649	135	39	280	84

Table 3.1 Table of data sources for PWM data. Shown are the total number of peak in all factors for each source, the number of distinct factors, the number of distinct timepoints/cell lines for each source, the number of datasets, and the number of datasets with matching PWMs.

Source	PWMs	Factors	With ChIP Data
flyfactor	670	359	64
jaspar	131	126	40
pouya	200	62	60
berkeley	5	5	5
berkeley bdtnp	19	19	19

Table 3.2 Table of data sources for PWM data. Shown are the various databases from which PWMs were gathered, the number of Position Weight Matrices from each, the number of distinct factors, and the number of distinct factors with information in the ChIP database.

3.1.1 Receiver operating characteristic analysis of PWMs

A PWM is a matrix that is used to assign a score to a given DNA, RNA or protein sequence, by assigning weights to each nucleotide (or amino acid) at each position (for some number of positions, usually 4-30 bp) and then summing the weights for each base (Stormo *et al* 2000). When used to express a TF motif, such a matrix can

(though need not) be interpreted as the log-odds score of finding a given base at a given position in a binding site, relative to the genomic background. The higher the score assigned a given piece of DNA, the higher the likelihood that a piece of DNA represents a binding site for the factor. PWM scores encode the assumption that bases affect affinity independently, which is an approximation, but one which has still allowed PWMs to become the de facto standard for modeling protein binding. Many PWMs are derived from *in silico* motif discovery tools, which have a strong tendency to over-fit their input data (Simcha *et al* 2012). Often the motifs they output, although statistically enriched in the input data, lack predictive power, particularly when used with other ChIP datasets – for instance in a different cell type or developmental stage, or even a different antibody. Motif discovery tools will often yield motifs that simply reflect the sequence biases of the input data – e.g. motifs composed entirely of di or mononucleotide repeats (Fig 3.1a). In addition, they may yield motifs that actually represent binding factors for TFs other than the one used to generate the ChIP data (Fig 3.1b). We therefore decided to carry out receiver operating characteristic (ROC) analysis on our PWM motifs. A ROC curve is a tool used to evaluate the performance of some binary classifier (such as a PWM) as its discrimination threshold is varied. Concretely, it is the curve formed by graphing the true positive rate against the false positive rate. The AUC (area under curve) is the fraction of the graph's total area that falls underneath this curve, and is a commonly used metric to assess the quality of a binary classifier. Motifs discovered experimentally, for instance by protein binding microarrays by (PBMs) are likely to better reflect the biochemistry of the factor when binding DNA in vitro, however this may not reflect behavior *in vivo* .

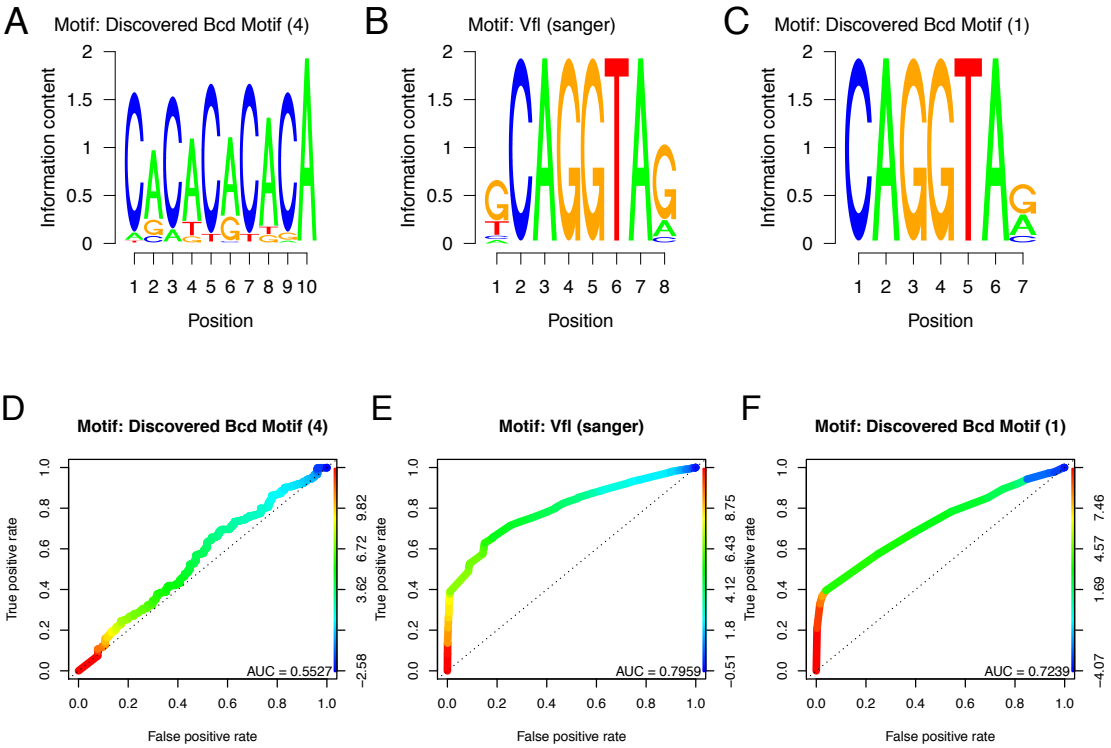


Figure 3.1 Examples of discovered motifs, with ROC curves.

A) A dinucleotide repeat motif identified as enriched within Bcd (Bicoid) ChIP-chip peaks by Meme. Such low complexity repeats are a ubiquitous feature of enhancers, and are often identified by motif discovery tools. They typically lack predictive power (D) as evidence by their low AUC scores when using DNase sensitive regions as negative controls. B) The canonical PWM for the transcription factor Vfl (also known as Zld) is highly predictive (E) and variants of this motif can be found as enriched in the ChIP peaks of many other factors – e.g. C) shows a motif discovered in the chip peaks of Bcd. This reflects the actual biology of the factors – the two motifs co-bind many enhancers – but represent an obstacle when linking ChIP data to motif data. F) Shows the ROC curve for the Zld motif when predicting Bcd binding, note that the Zld motif predicts Bcd binding as well, though not as well as the binding of Zld itself (E).

We therefore examined the receiver operating characteristics of each of our transcription factor motifs, using the highest quality available datasets as positive sets, and a published embryonic DNase sensitivity peaks, without the relevant ChIP dataset, as negative regions. The choice of DNase sensitive regions as negative regions is important, as the negative set should have sequence characteristics as close as possible to the positive. Some factors showed extremely high AUCs, for instance, one motif for Zld (also known as vfl) – a factor involved in the activation of the zygotic genome during the maternal to zygotic transition, shows an AUC of 0.80

(Fig 3.2e). This likely reflects the fact that Zelda serves as a 'pioneer' factor (Sun *et al* 2015), and is therefore able to bind even closed chromatin, without the aid of other factors, so that sequence affinity becomes the primary determinant of its binding. Conversely, some factors actually showed negative AUC, meaning that their presence anti-correlated with occupancy for their supposed factor across DNase sensitive sites. These likely represent instances of motifs that have been attributed to an incorrect parent factor. Overall, the distribution of AUCs for our motifs (Fig 3.3) was highly variable, and we identified 398 motifs with an AUC of 0.6 or more, and applying this value as a filter, we used these motifs in subsequent analysis.

3.1.1 Selecting motif thresholds

One difficulty that arises in studies that make use of PWMs is that PWMs yield a quantitative score, rather than a qualitative bound/non-bound result. There is no universally correct way of threshholding PWM scores, and so studies (e.g. Ohler *et al* 2002, Negre *et al* 2009) have chosen to sidestep this issue by simply setting an arbitrary cutoff on its p-value, such as 1e-4, where this p-value is derived by some low-order markov model of sequence probabilities. This approach essentially requires that TF binding sites are rare. As such it will work reasonably well for motifs with fairly specific sequence affinities, but poorly for motifs that are in fact rather nonspecific. Another approach used by the program Patser (Stormo *et al* 1999) uses the information content of the motif itself to set the score cutoff, and involves the assumption that the PWM is an accurate description of the factors binding affinity for DNA. This approach is useful in that, given PWMs that accurately describe sequence affinity, it will accurately call many sites for nonspecific factors and few for highly specific factors. In practice however, the score will be at best a very rough guide for protein binding, since it will fail to account for the many other interactions with nucleosomes, other factors, molecular crowding etc. Motifs produced by a statistical motif discovery program such as Meme, will also contain only incomplete information, compared to a PWM generated by a PBM. Biochemical affinity is in any case often too complicated to express as a linear weighting over nucleotides. This method therefore often gives poor results, with extremely low specificities when compared to actual datasets.

In order to avoid the biases inherent to these methods, were therefore chose to use 50% sensitivity as an arbitrary score cutoff at which to examine the properties of TF binding sites. Previous work (Spivakov *et al* 2012) suggests this is a reasonable cutoff for many factors. We reasoned that factors whose actual bound motifs are found at a much higher frequency within their ChIP binding sites would see only a fraction of these sites excluded by this method, and those with a much lower fraction of actually bound sites, which would likely show little signal in further analysis in any case, would be excluded by our subsequent filter.

3.1.2 Examining Motif Enrichments

In addition to analyzing AUC for each motif within a DNase sensitive background, I carried out a second analysis of Motif quality by assessing the genomic enrichment of motifs within their ChIP peaks, divided by the enrichment of shuffled control motifs (Fig 3.2). (See Material and methods). This analysis complements the AUC based analysis in that it is not reliant on DNase peak data, and takes into account the genome as a whole. In addition, the use of shuffled motifs as a comparison controls for predictive value the PWM may have due to its GC content alone, a factor which may be more or less common for certain PWMs, and for certain ChIP datasets. Finally, the enrichment score is a measure that specifically assesses the performance of the cutoff selected for the PWM, rather than its performance over the range of cutoffs selected. The measure is also more conservative than simply taking the enrichment of non-shuffled motifs alone. We identified a total of 309 Motifs with enrichment of 1.1 or more in their target ChIP peaks, and applied this as a second filter to our data.

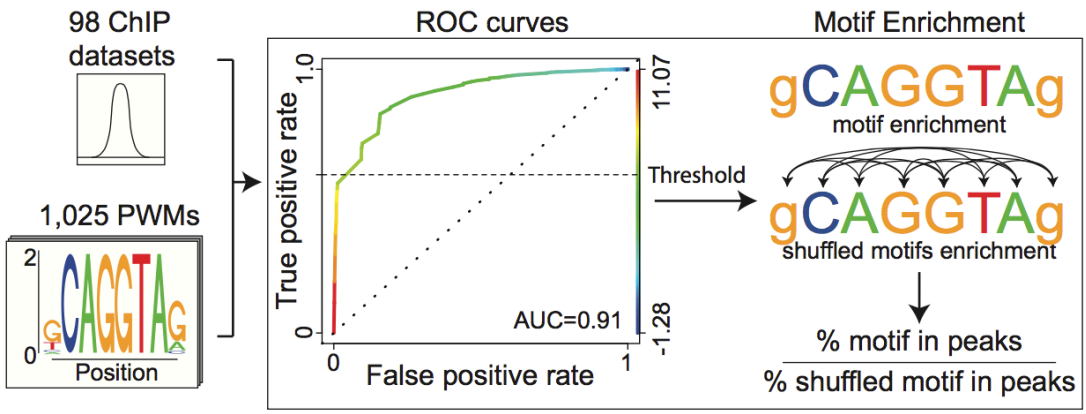


Figure 3.2 Diagram showing procedure for filtering TF motifs.

Shown is a schematic representation of the procedure for deriving AUCs, and enrichments cores. PWMs were first intersected with their relevant ChIP datasets, and whole embryo DNase sensitivity data. DNase peaks bound or unbound by the relevant factor were used as true positives and negatives respectively, to derive an AUC for each motif. The genome wide fraction of the motif's hits inside its ChIP peaks was then divided by the genome wide fraction of shuffled control-motif-hits inside

3.1.3 Correctly assigning motifs to their target factor

One common issue in motif discovery is that motifs enriched within the ChIP peaks for a factor may not be the motifs bound directly by that factor, but rather, motifs bound by other, co-binding motifs (Fig 3.1 a, b). We attempted to resolve this issue by first grouping similar motifs into clusters, using their normalized column-wise correlation score (Pietrokowski 1996). We then took only motifs which had the highest enrichment for their target gene within each cluster, and which did not have a higher enrichment in the peaks of the target gene for a motif in the same cluster. This yielded a total of 274 non-redundant motifs. Of these, 78 PWMs for 36 distinct factors were also 'High Quality' with respect to AUC and Enrichment, and we selected them for further analysis. (Fig 3.3).

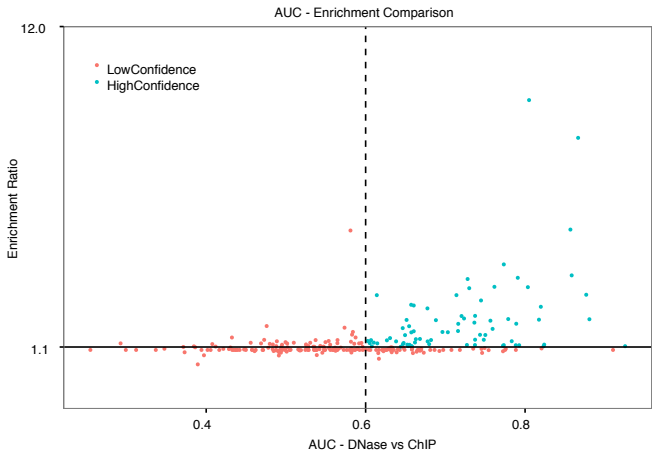


Figure 3.3: AUC vs. enrichment score for TF motifs

Enrichments score – i.e. the proportion of the motifs found in ChIP peaks for their factor, divided by the fraction of control motifs in the relevant factor (y-axis) vs. AUC – defined as the maximum AUC for each motif as a predictor of which DNase sensitive regions are bound by its factor. We defined high confidence TF motifs as those with and AUC (x-axis) of 0.6 or more, and an enrichment score (y-axis) of 1.1 or more. The measures are correlated, since both give a read out of quality, but are nevertheless somewhat independent means of assessing PWMs.

3.1.4 Functional analysis of TF Motifs

Having applied our PWM quality assessment pipeline to our TF motifs, we wished to use a metric independent of ChIP data to assess the accuracy of our pipeline, a task for which INSIGHT is suited (Gronau *et al* 2013), since it does not rely on ChIP-seq or similar annotations. Since our analysis of eQTL and INSIGHT are both methods that derive information from allele frequencies, filtering our motifs using INSIGHT, and then carrying out eQTL analysis, would be somewhat circular. We could therefore not use INSIGHT to further filter motifs, but could use it as an independent confirmation of our pipeline's validity. INSIGHT can provide an estimate of various parameters related to selection, including rho – the proportion of functional sites in the regions of interest. This estimate accounts for demographic influences as well as differences in mutation rates and coalescent times across the genome.

In order to control for biases in functional base content due to CG content, differential quality in ChIP datasets etc., we devised a score – Rho-Enrichment –

analogous to the enrichment score based on presence in ChIP peaks. Rho-Enrichment is generated by dividing rho – the fraction of functional sites – in actual motif matches by the value of rho in shuffled motif matches. We filter both for presence in the relevant set of ChIP peaks. This provides an independent measure of the motifs biological functionality. It is likely however to be a somewhat noisy measure of functionality. Because it compares regions within ChIP peaks, the inclusion of many 'false positive' in the ChIP dataset, or the inclusion of a large amount of non-regulatory sequence (for instance due to the imprecise nature of ChIP-chip) will tend to decrease its power. Furthermore there can be no guarantee that the non-shuffled motifs will not resemble other functional motifs, or simply overlap them since they are within peaks, and in cases where peaks are unusually narrow or a factor binds with other motifs unusually often, this will also make Rho Ratio tend towards 1. The measure also requires that enough peaks and motifs for the motif be present, in order to accurately estimate rho values.

Figure 3.4a shows the rho values obtained for a selection of our TFs. Of our 274 non- redundant motifs, 125 had a significantly higher fraction of functional bases than the corresponding shuffled motifs (where significance was assessed using INSIGHT's estimates of standard error for the rho parameter). We also used INSIGHT to ask whether our motifs had more adaptive selection that their shuffled counterparts (3.4b). Indeed, 56 (5%) of our motifs showed evidence for significantly greater adaptive substitution than shuffled control motifs. There was no significant correlation between the ratio of adaptive substitution between motifs and controls and Rhoratio. High quality motifs were not more likely to show greater adaptive substitution than control motifs.

Figure 3.5 shows the relationship between two metrics, AUC and enrichment score, and the Rho Ratio. Both correlate with Rho Enrichment, indicating that these two metrics correlate with motif's enrichment for sequence conservation and adaptation. Both provide evidence for biological function, and hence, these results provide independent evidence that our motif-filtering pipeline is biologically meaningful. Our 'High quality' motifs were significantly enriched for functional motifs as assed by INSIGHT (odds ratio 95% CI 3.99 – 15.4, $p < 3.86 \times 10^{-12}$, fishers exact test) with 79% of the high Quality, and 32 % of the low quality motifs, showing

a significant Rho ratio. We furthermore used a linear model to ask if the two measurements correlated independently with Rho Ratio and found that indeed, both have independent associations with Rho Enrichment, validating our use of both measurements to filter our motifs (see methods). For a small minority of TF motifs, Rho Ratio is negative, however these motifs (e.g. the flyfactor motif for ubx) in general have a lower absolute value of rho, and none pass our AUC/Enrichment based quality filter, suggesting that they are artifacts of the motif discovery process. In such cases, shuffled control motifs may well contain motifs that are in fact functional, which could results in negative Rho Ratios.

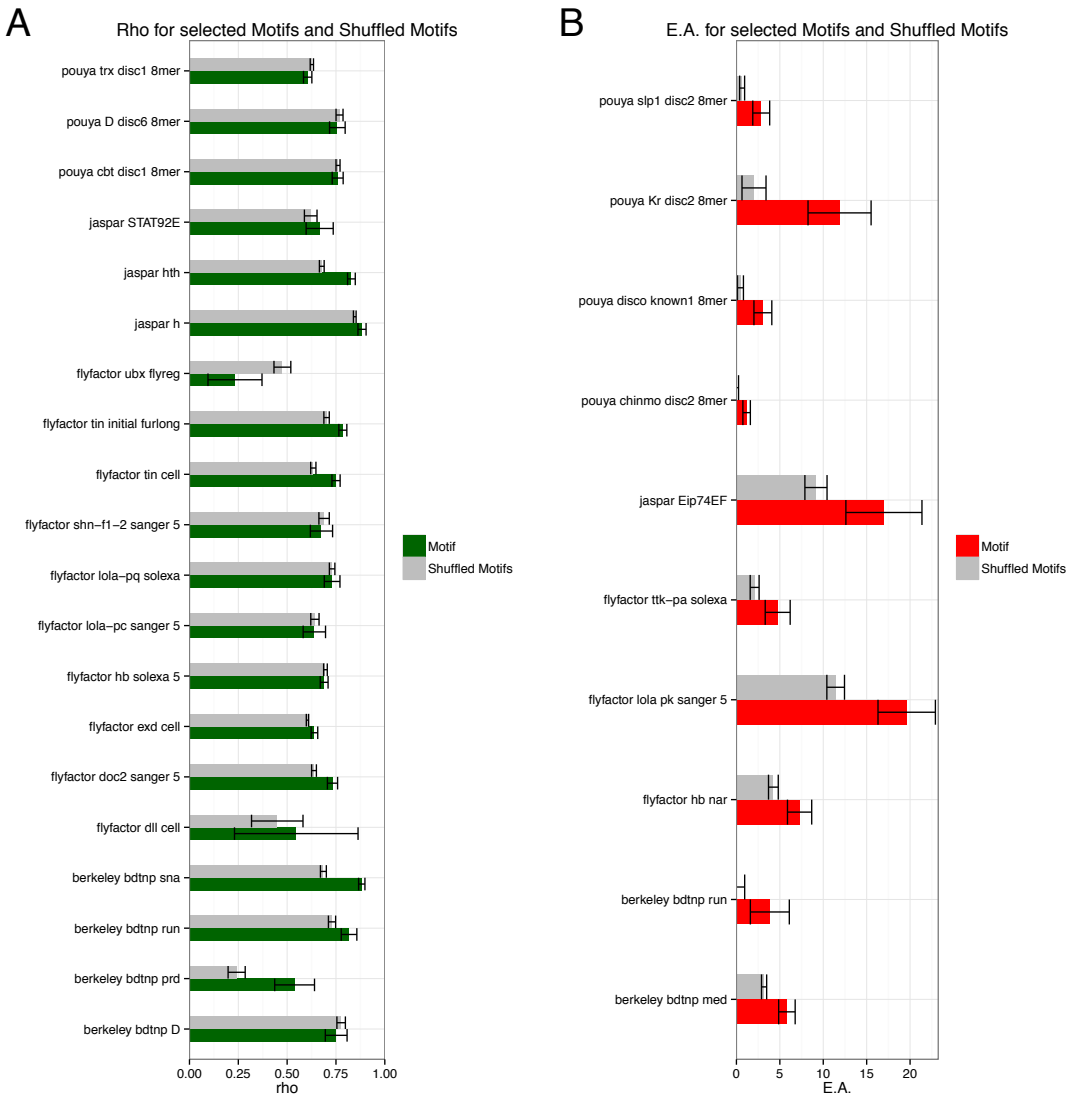


Figure 3.4: Negative and Positive and selection on transcription factor binding sites.

A) Rho (estimated proportion of functional base pairs) for 20 TF motifs and shuffled versions of them**.** Transcription factors are selected evenly from across the distribution of Rho values. Both shuffled and non-shuffled motif matches overlap peaks of the relevant ChIP dataset. B) Estimated adaptive substitutions per kilobase for the 10 transcription factors with the most adaption relative to shuffled control motifs.

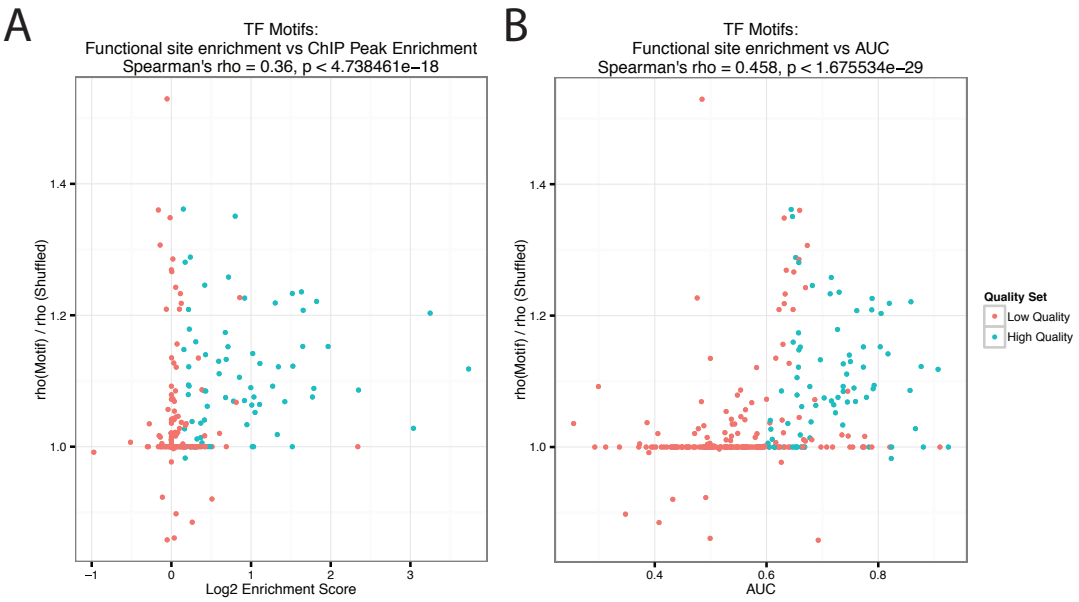


Figure 3.5: Comparison of Rho Ratio to AUC and enrichment score.

A) Relationship between Rho Ratio and enrichment score. Rho (estimated proportion of functional bases) was estimated for motifs and for theirs shuffled controls, and the conservative estimate of the ratio between these was plotted on the y-axis. The x-axis shows the log2 enrichment score – i.e. the proportion of the motifs found in ChIP peaks for their factor, divided by the fraction of control motifs in the relevant factor. B) Same y-axis as A), but the x-axis shows the AUC for each motif. Both Enrichment (A) and AUC (B) (x-axis) show a significant association with the enrichment of functional base pairs in the motif over controls (y-axis).

3.2 Discovering promoter-associated motifs in *D. melanogaster*

In addition to my study of transcription factor motifs in *D. melanogaster,* I also carried out an analysis of promoter-associated motifs. Previous studies (Ohler *et al* 2002, Fitzgeral *et al* 2006,Down *et al* 2007)have made use of more limited sets of *Drosophila* TSS, and as such my set of CAGE peaks offers an opportunity to discover previously unknown promoter associated motifs, and variants of known motifs. Furthermore, we wished to base our mechanistic analyis of QTL on motifs which were particularly abundant within our CAGE peaks, and thus we elected to carry out our own motif discovery. To discover these motifs I made use of the Meme Suite (Bailey *et al* 2009), which is particulary suited to the analysis of large sequence datasets, and also offers analysis of positional enrichment.

For sequence analysis, I defined three sets of sequences – 'Narrow Promoters '(shape index > -1), 'Broad Promoters' (shape index < -1, and 'all promoters'. These sequences were all drawn from the set of 'main' peaks (see previous chapter). 'Internal' peaks were also analyzed with Meme, however this analysis did not yield any motifs resembling known promoter associated motifs, instead yielding only some low confidence motifs resembling known splice signals, supporting our choice to exclude internal peak from the main analysis, and suggesting that the internal peaks frequently represent a different biological phenomenon, such as recapping or contamination in the CAGE protocol, than the 'main' set.

We identified promoter associated motifs in *D. melanogaster* by analyzing the broad and narrow categories individually, the set of all promoters together, and also by looking for motifs enriched in broad peaks against the background of narrow peaks, or vice versa. Each of these runs yielded a set of motifs, 181 in all (Table 3.3 - appendix). Since many of these motifs are highly similar to one another (for instance the INR motif appears when examining all TSS, but also in two different variants when contrasting narrow vs. broad promoters, and three different variants when examining only narrow TSS). We therefore designed a similarity score that combines similarity of sequence, and similarity of positional enrichment, to cluster our promoter associated motifs. The similarity score is defined as the column wise

correlation score used by Pietrovski *et al* (1996), with 0.1 added for motifs whose zone of positional enrichment, as determined by Centrimo, overlapped by 50% of more. Clustering with this metric yielded a total of 58 clusters.

Source	PWMs	Factors	With ChIP Data
flyfactor	670	359	64
jaspar	131	126	40
pouya	200	62	60
berkeley	5	5	5
berkeley bdtnp	19	19	19

Table 3.3 Table of discovered promoter associated motifs.

Shown are the various databases from which PWMs were gathered, the number of Position Weight Matrices from each, the number of distinct factors, and the number of distinct factors with information in the ChIP database.

Each of the eight most studied promoter motifs in *Drosophila melanogaster* (Ohler *et al* 2002, Ni *et al* 2010, hereafter referred to as the 'Ohler Motifs') cluster with one or more identified motifs in our set, indicating that our motif discovery procedure recovers the known sequence features of *Drosophila* promoters. All 15 enriched k-mers identified by Fitzgerald *et al* (2006) are also encompassed in one of our clusters, as are 50% of the 120 motifs identified by Bailey *et al* (2009), which likely represent a larger, less stringent set of motifs, with more false positives.

Of our 59 clusters of similar motifs, 65% include a motif with significant positional enrichment, (Fig 3.6). These positional enrichments vary in both breadth and location. For instance, motifs that cluster with the INR motif (Fig 3.6a), as expected, are tightly positioned around the transcription start sight, as are motifs clustering with Motif 1. Motifs matching the TATA box show enrichment upstream of the TSS, while motifs matching DPE (downstream promoter element) are enriched downstream of the TSS. Of the motifs that do not match any of the eight Ohler motifs, many also show strong positional enrichment (Fig 3.6b). For instance, the motif AAYSGAA, identified in narrow promoters, is strongly centrally enriched

around TSS. Meanwhile several low complexity motifs resembling dinucleotide repeats show enrichment ~ 220bp upstream of TSS.

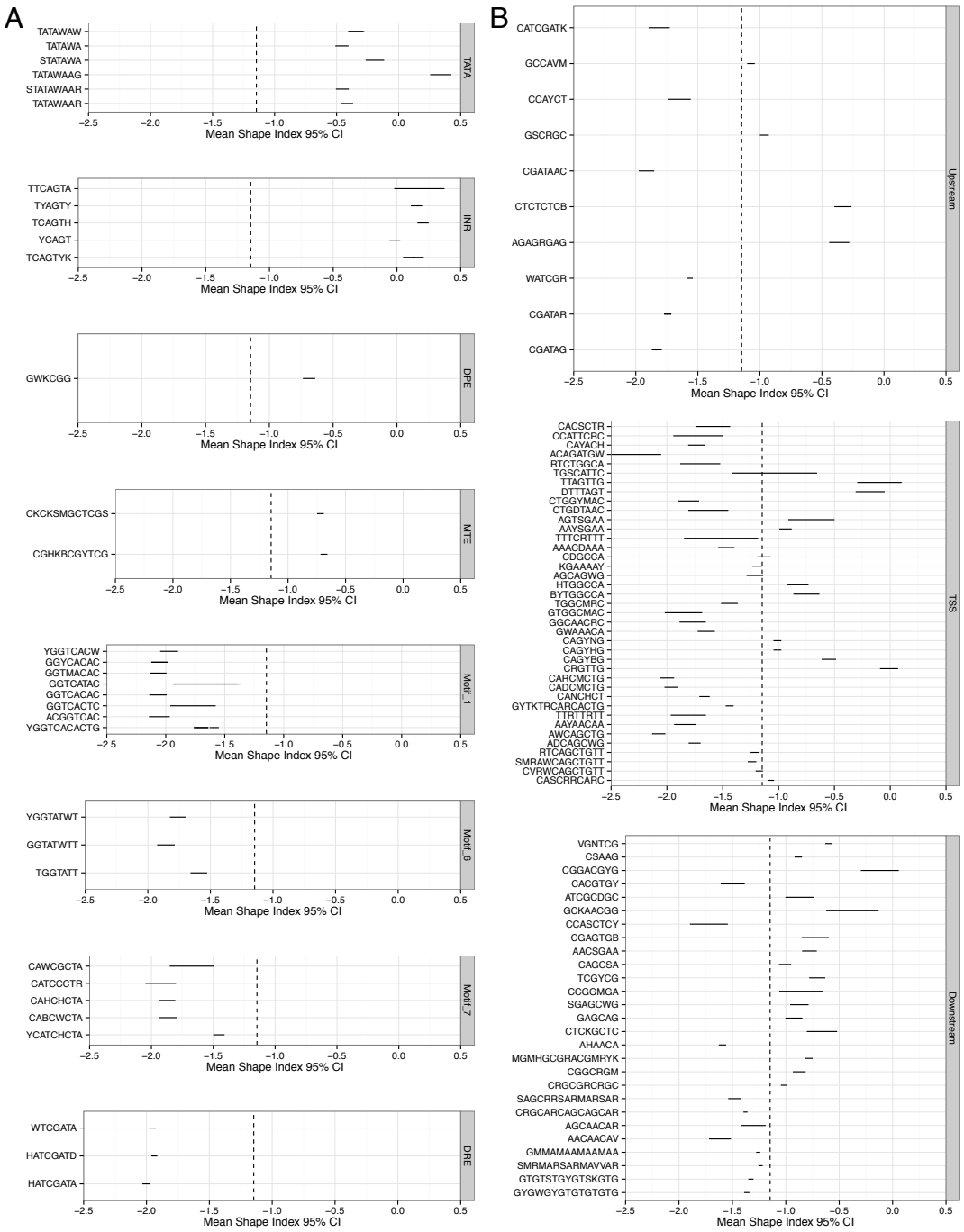


Figure 3.6: Positional enrichment for promoter associated motifs

A) Positional Enrichment for Discovered Promoter associated motifs resembling Ohler Motifs. Motifs similar in sequence and positional enrichment to the known Ohler motifs (panel labels) were recovered by our motif discovery procedure. Dotted

lines represent the +/- 10 bp zone surrounding the TSS, dividing upstream, TSS, and downstream motifs in our subsequent classification. The x- axis shows the zone of enrichment defined by Centrimo, with the dot denoting its centre. B) Positional Enrichment for Discovered Promoter associated motifs. We divided motifs not matching a known Ohler motif up into upstream, downstream, and TSS associated motifs.

Our discovered motifs also vary widely in the number of instances at which they are found in promoters (Fig 3.7). Our TATA-like motifs range in prevalence from ~4% to ~22%, depending on their specificity. This is in broad agreement with the numbers for TATA promoters usually quoted in the literature. We note that there has been some dispute (Fitzgerald *et al* 2006) over the prevalence of the TATA box as a promoter associated motif in the literature. The apparent prevalence of a motif is strongly dependent on the background model used for random sequence (AT rich sequences like the TATA box will be found dramatically more often if CG rich regulatory regions are used as the background) and the specificity of the motif, as well as the area relative to the TSS in which instances are counted (we used the positionally enriched area for each of our motifs) we suggest this is the reason for the disagreeing reports. Motifs not matching Ohler motifs also show a wide range of prevalence, from several short motifs and AT rich repeats that are present in over 50% of promoters, to motifs that are found in less than 5% of promoters. Also highly variable is the degree to which promoter motifs are enriched over their expected abundance based on base composition. While promoter motifs like INR, DRE and TATA show a dramatically higher prevalence than their shuffled counterparts, many of our more highly prevalent promoter-associated motifs are only somewhat enriched over shuffled control motifs with the same base composition. This may indicate that their presence is simply a result of Meme's failure to capture more complex sequence biases in background sequences. In conclusion, because our study of promoter associated motifs uses a larger input promoter set than other similar studies in *Drosophila melanogaster*, we have been able to identify more rare motifs that may play a role in only a small percentage of promoters.

Figure 3.7 Promoter associated motif prevelance

A) Fraction of 'main' CAGE peaks containing discovered motifs (x-axis) resembling Ohler Motifs, and same for shuffled control motifs. Each motif's consensus (y-axis) and its matching Ohler motif (grey box) are shown. Only motifs within the zone of enrichment defined by Centrimo were counted. B) Same as (A) for promoter associated motifs not resembling Ohler Motifs. Discovered promoter associated motifs vary widely in their prevalence, and enrichment over shuffled controls, with some shorter, simper motifs being present at ~50% of all TSS, and others being present at less than 1%. The motif's location class (grey box) is shown.

We then wished to examine whether any of our newly discovered promoter associated motifs showed strong preferences for broad or narrow promoters, in the same way that the known Ohler motifs do. We plotted the mean shape index for CAGE peaks containing each of our motifs, along with 95% confidence intervals for this value obtained by bootstrapping. We find that, as expected, motifs resembling known Ohler motifs show bias in their shape index (Fig 3.8a), with motifs resembling INR and the TATA box for instance, showing a high average shape index, indicating they are found more often in narrow promoters, while motifs like Motif 1 and DRE found more often in broad promoters. The rest of our motifs (Fig 3.8b also show significant shape biases. For instance the motif AAYAACAA, which resembles one motif (TIFDMEM0000065) discovered by Down *et al*, shows a significant (cohen's d = 0.38 – 0.54, 95% CI, Welsh's two sample t-test) bias towards broad promoters. This motif also resembles the motif for the zinc finger domain TF CG4360, a weak ortholog of vertebrate myoneurin, whose function is currently unknown. We note that although motifs discovered by contrasting broad and narrow promoters can be expected to show shape bias (as indeed 97% do), 78% of motifs discovered in the set of all TSS also show significant (Welsh's two sample t-test, $p < 0.05$ after correction for multiple testing) bias towards narrow or broad promoters, emphasizing the distinct characteristics of regulation at these classes of promoter.

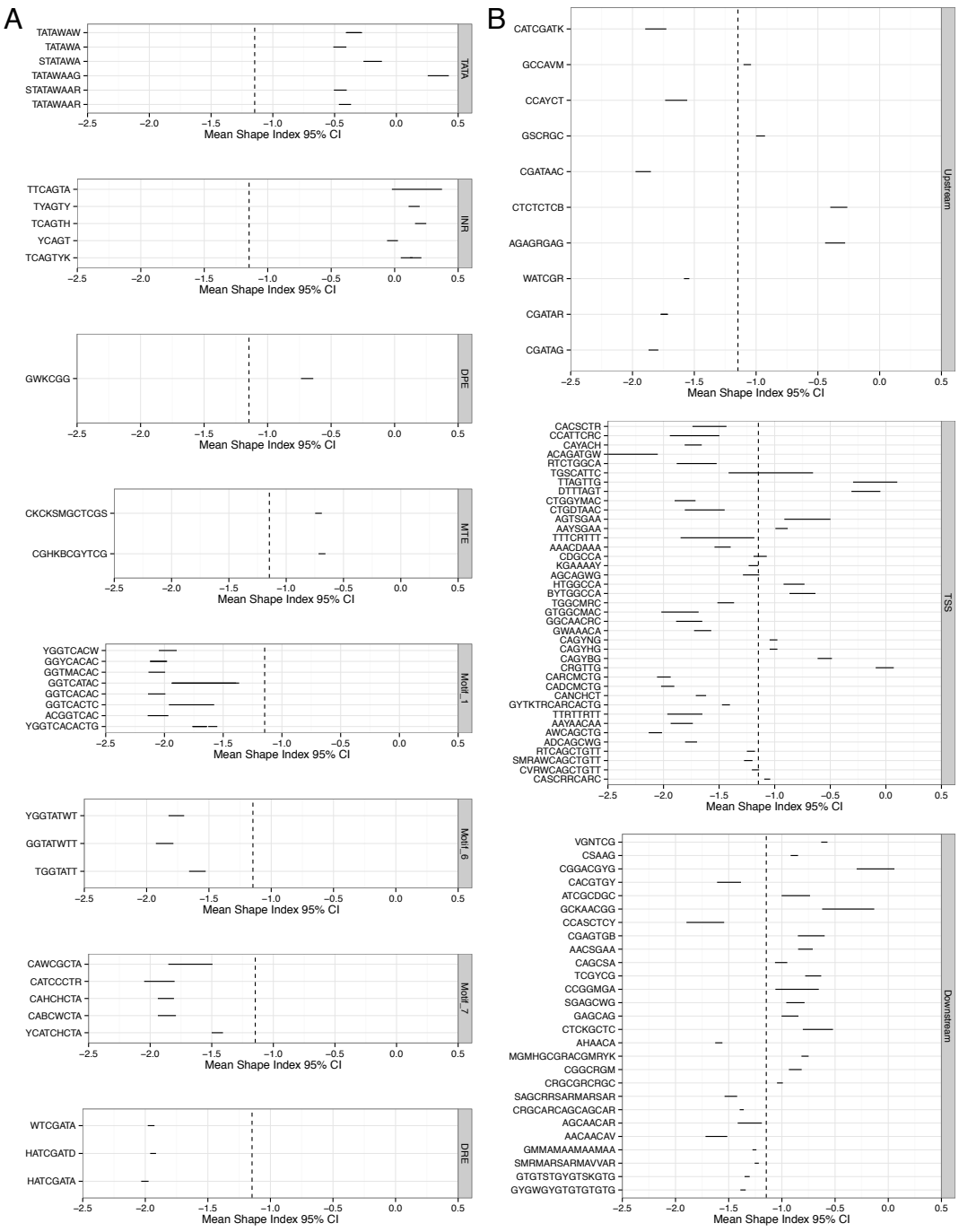


Figure 3.8 Shape bias of discovered promoter associated motifs

A) Shape bias of Discovered promoter associated motifs resembling Ohler Motifs. TSS with discovered motifs which match Ohler motifs have mean shape scores differing from the global shape score mean, reflecting the known biases of the Ohler motifs (e.g. INR is biased towards narrow promoters). Bars (x-axis) represent 95% CI limits of the mean shape index for all CAGE peaks with the motif. Dotted line represents average shape index of all CAGE peaks.B) Shape bias of discovered promoter motifs not resembling Ohler Motifs. TSS with discovered motifs that do not match Ohler motifs also have mean shape scores differing from the global shape score mean.

3.2.1 Functional analysis of discovered promoter-associated motifs

Having identified motifs enriched in *D. melanogaster* TSS, we then wished to apply our function analysis framework to the promoter-associated motifs as well. Unlike transcription factor motifs, our promoter-associated motifs lacked any specific sequence-independent measure that could provide a region in which functional motifs would be present. As a proxy, we instead used the zones of positional enrichment for each motif, as defined using Centrimo, relative to the 'main' set of TSS. Figure 3.9 shows the rho values for our promoter motifs alongside their shuffled counterparts. We observe that many of our promoter-associated motifs show a high Rho Ratio, including all those motifs that match the eight well-studied Ohler motifs (Fig 3.9a,b). Many of our three positionally enriched unknown classes also show a significant Rho Ratio. We also compared the proportion of motifs with significant Rho Ratios in the 'Upstream', 'TSS' and 'Downstream' (Fig 3.10). This analysis indicated that, surprisingly, TSS motifs were the least likely to show enrichment in functional bases over their shuffled versions, and upstream motifs were the most likely. A likely explanation for this is the regions used for the comparison – TSS are in general highly enriched for functional bases, such that many of the shuffled motifs would still overlap functional regions. Similarly, downstream regions being transcribed are more likely to be functional.

In all these results provide further evidence that our promoter-associated motifs represent biologically meaningful sequence motifs, and suitable basis on which to examine the mechanisms of variation at transcription start sites.

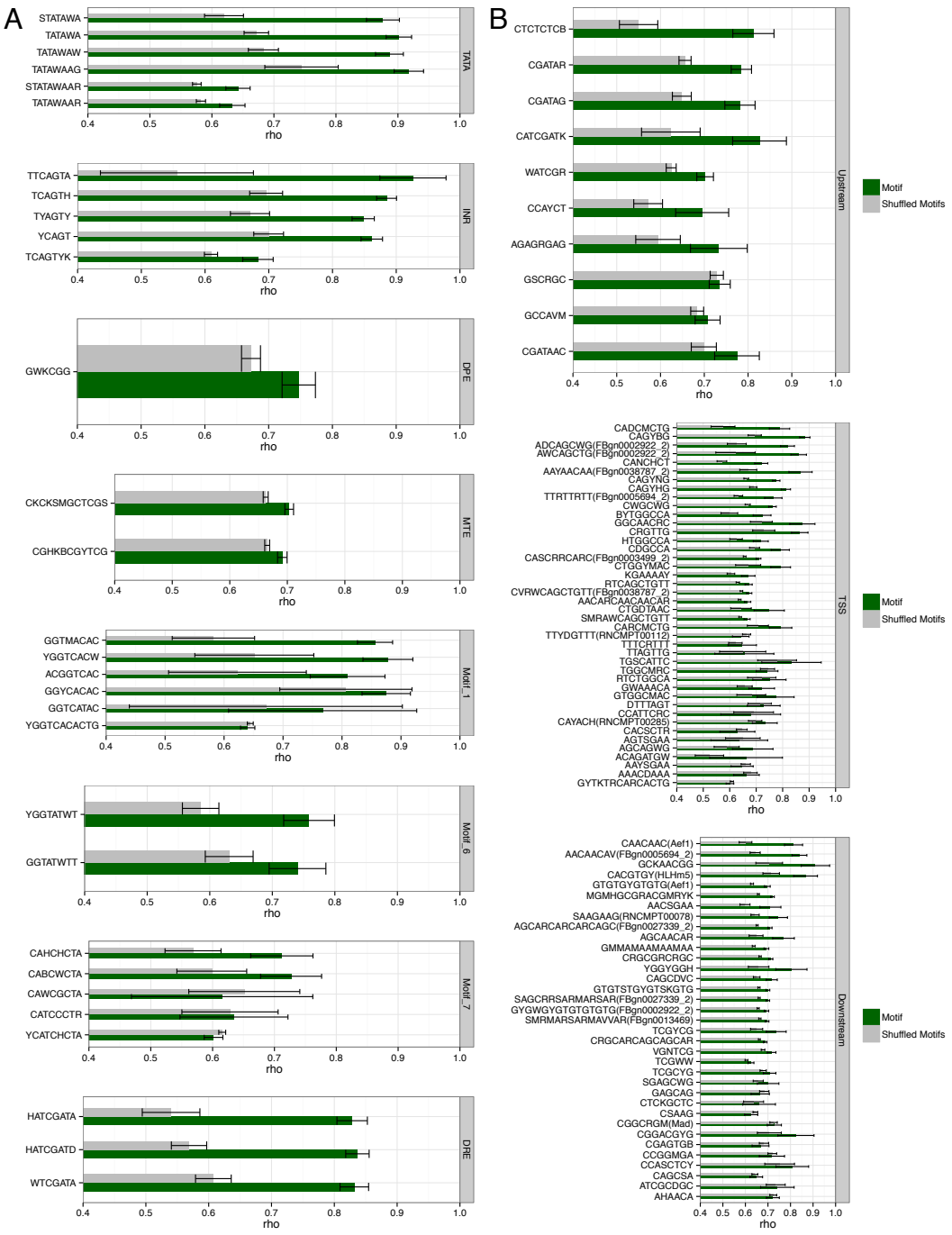


Figure 3.9 Shape bias of discovered promoter associated motifs

A) Rho values (x-axis) of discovered promoter-associated motifs resembling Ohler Motifs (panel labels). Ohler motifs re-discovered from CAGE peaks often show enrichment relative to shuffled comparison motifs, in some cases (such as DRE and the TATA box) quite strongly. Error bars represent the standard error of the parameter estimate from INSIGHT. B) Rho values of discovered promoter-associated motifs not resembling Ohler Motifs. Motifs discovered from CAGE peaks often show enrichment relative to shuffled comparison motifs, with upstream motifs in particular often showing enrichment for functional bases.

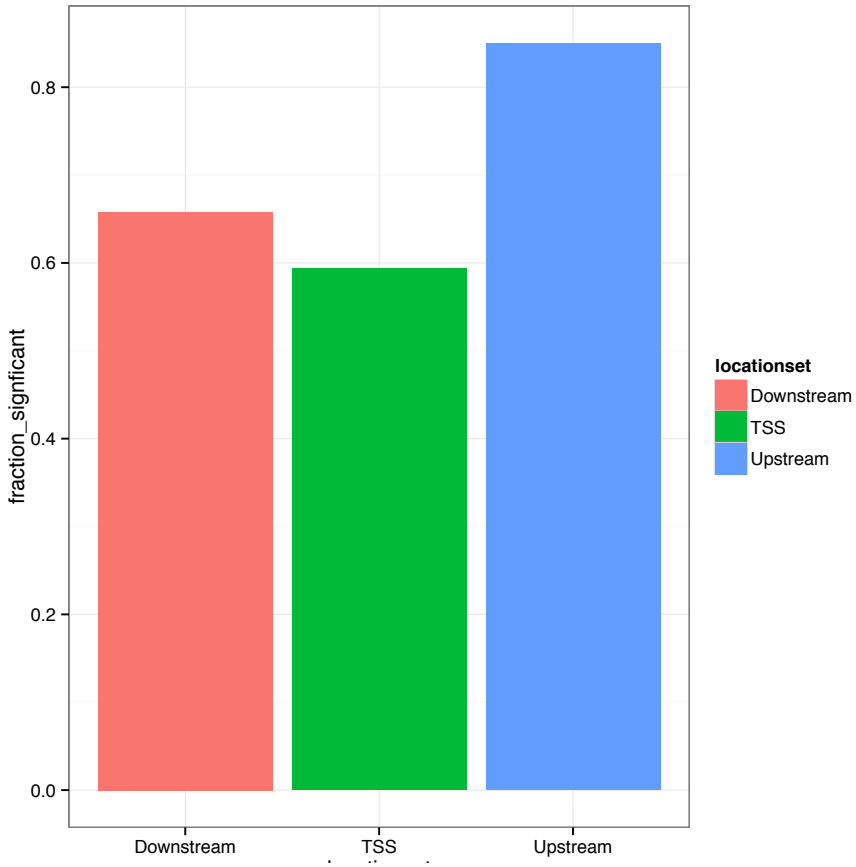


Figure 3.10 Proportion of promoter associated motifs with significant Rho ratio.

Fraction of motifs whose Rho value differs (as per 95% CI calculated using INSIGHT's standard errors, y-axis) from shuffled comparison motifs for three classes of promoter-associated motifs – a significantly (Fisher's exact test, $p < 0.05$) higher proportion of the upstream motifs are enriched for functional base pairs.

3.3 pA site-associated motifs in *D. melanogaster*

3.3.1 Discovering pA site associated motifs in *D. melanogaster*

In addition to my study of Transcription Factor Motifs in *D. melanogaster,* I also wished to study motifs found in proximity to pA sites. As with the discovery of promoter-associated motifs, we faced *D. melanogaster* the lack of any specific genome wide read assay corresponding to individual motifs – a given RNA motif could be present in a small percentage of, or almost all, polyadenylation sites, meaning that ROC analysis of our motifs was impossible. We therefore decided to use a simple consensus sequence based approach for our polyadenylation motifs – an approach that has been used in previous studies for RNA motifs (Ray *et al* 2013). We again made use of the Meme Suite to discover pA site associated motifs in *Drosophila* (see materials and methods). As our search sequence, we used the 300bp, centered on the 3' end of each 3' Tag-seq peak. We chose an E-value cutoff of 0.00005 for our analysis.

Polyadenylation occurs through the cleavage of RNA at specific motifs followed by the addition of a poly-A tail by poly-A polymerase (reviewed in Proudfoot 2011). CPSF binds to the polyadenylation Site (PAS), a motif typically located 10-30 bases upstream of cleavage sites, while CstF binds around 25 bp downstream (Fig. 3.11a). Motifs resembling motifs for both these factors were identified by our de novo analysis (see Table 3.4) of which six resembled known polyadenylation factors. We designated these as 'known' motifs. Given that polyadenylation has been well studied in D *melanogaster*, we were also surprised to identify 18 other motifs. We designated the remaining motifs as 'Unknown proximal' – those motifs that did not resemble known motifs, but were positioned proximal to (within 25bp) of the cleavage site, 'Unknown distal' – which were enriched further than 25bp from the cleavage site, and 'Unknown Not Positioned' (see Fig 3.11b). Of these unknown motifs, some resemble known RNA binding proteins such as PABP and SUP-26. While the proteins binding the other motifs are unknown, their non random positioning, such as the AAACcaA motif positioned -1bp upstream of the cleavage site, and the TtCAtTT which is enriched 10bp downstream, suggests a specific function in the regulation of 3' RNA processing.

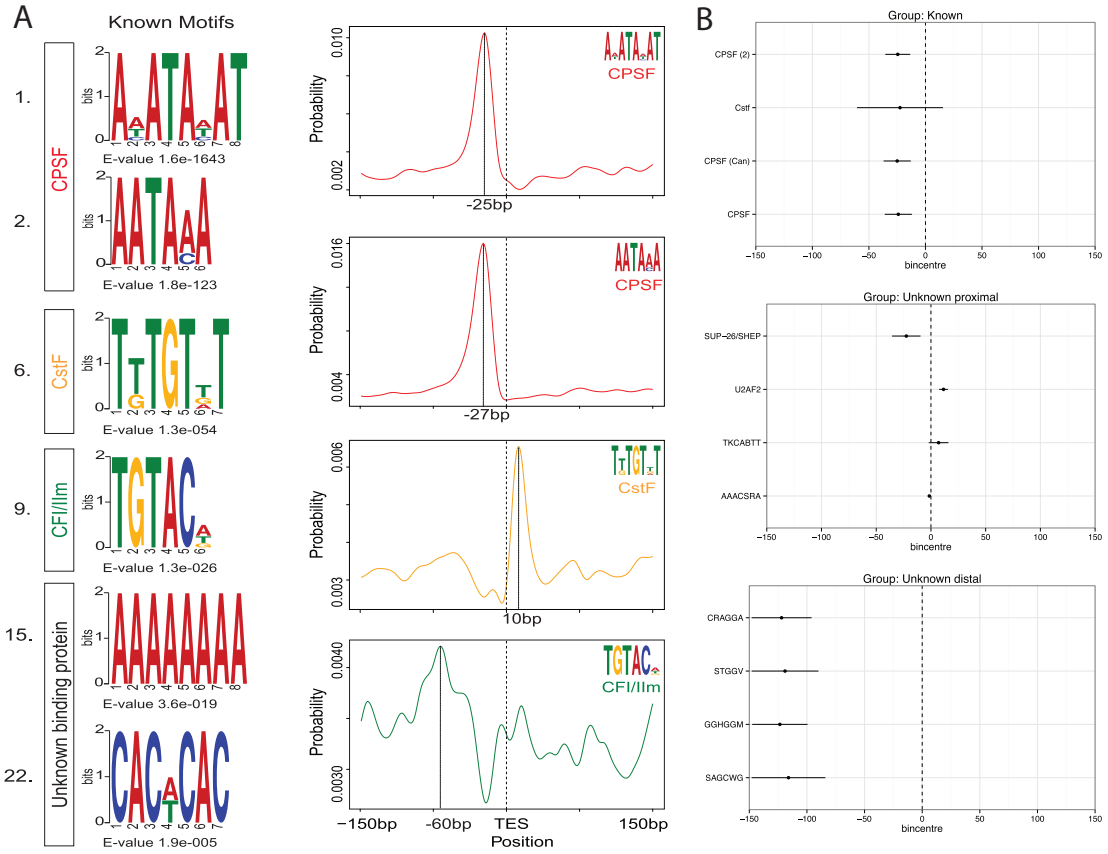


Figure 3.11 Discovering motifs associated with pA sites.

A) Consensus motifs and positional enrichment profiles for selected discovered pA site associated motifs. The known major polyadenylation motifs (1,2,6,9) in *Drosophila* were recovered, as well as other motifs (15,22). B) Positionally enriched motifs were divided into Known motifs, unknown but proximally enriched, and unknown but distally enriched. Shown are the zones of positional enrichment for each one, relative to the 3' end of the pA sites, which generally correspond to the strongest cleavage site for each pA site.

In accordance with their well studied role in *D melanogaster D melanogaster* biology, the six 'known' motifs are found at a higher prevalence than the unknown motifs, (Fig 3.12), with the canonical CPSF motif in particular being present at 38% of pA sites, in accordance with its role as the major cleavage motif in *D melanogaster*. If the other two variants of this motif are included, we find that 46% of pA sites show a CPSF motif, a figure that rises to 54% if only pA sites in the 'mRNA end' category are included. By contrast the unknown motifs are generally present in around 10% of TSS or less. All discovered motifs are enriched over shuffled control motifs, as expected, however this enrichment is often relatively modest, even for known functional motifs such as the canonical CPSF motif.

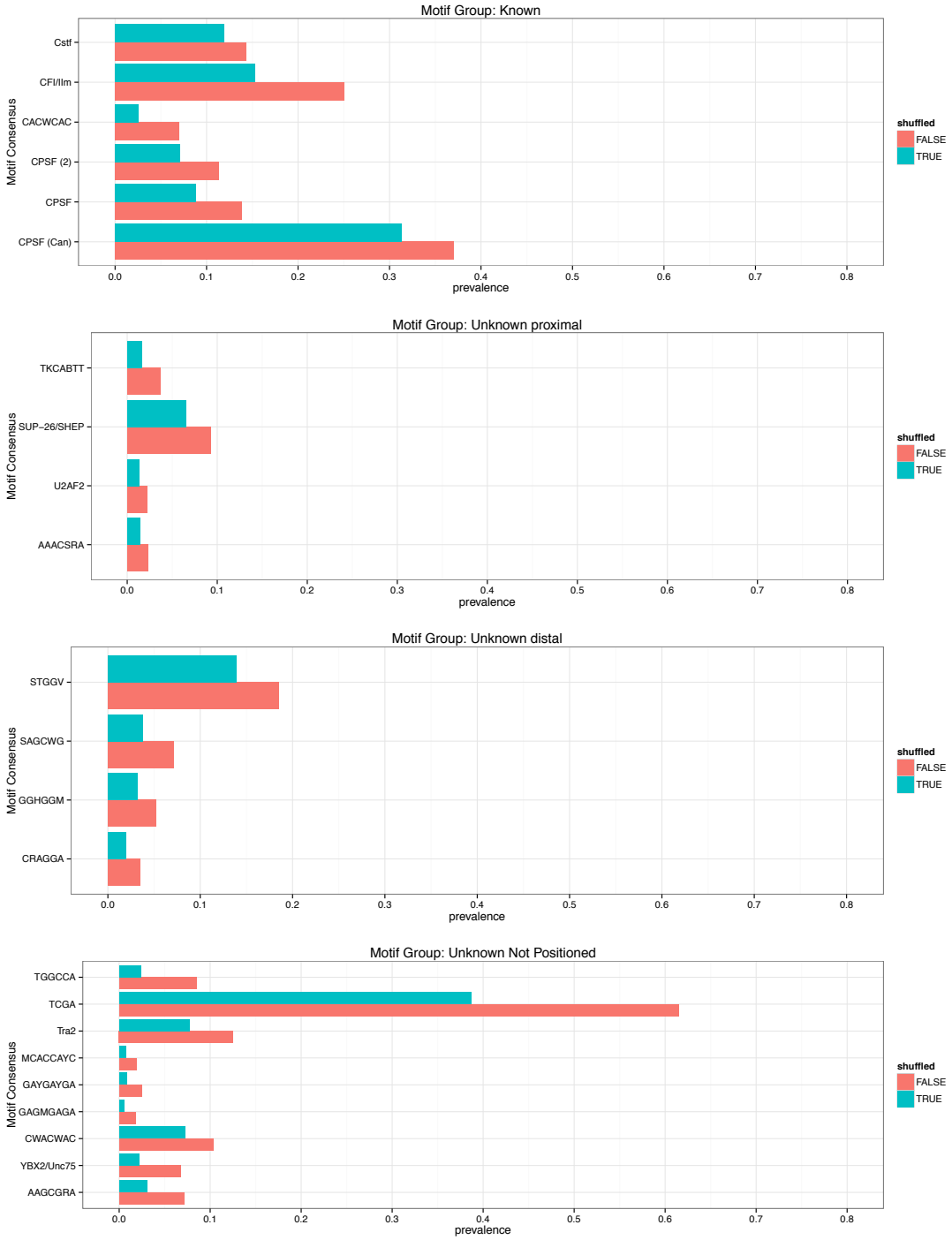


Figure 3.12 Frequency of motifs associated with pA sites.

pA site associated motifs (red) and their associated shuffled control motifs (blue) vary in the prevalence with which they are found in pA sites (x-axis). *De novo* discovered motifs (using all pA sites) were grouped into 4 classes; known (previously characterized pA motifs), unknown proximal (positioned within 30bp of the pA site, as per Centrimo) unknown distal (positioned further than 30bp from pA sites), and unknown non-positioned. The canonical CPSF motif is by far the most common known motif, at ~ 38%. Newly discovered pA associated motifs are in general less common than known ones, which may be the reason they have remained

uncharacterized. Note that in some cases motifs are only somewhat enriched over shuffled control motifs. Motifs were counted if in the zone of functional enrichment define by Centrimo.

3.3.2 Scans for Selection in discovered pA site associated motifs

Having discovered and categorized our pA site associated motifs, we then carried out functional analysis of them using INSIGHT (Fig 3.12a). We observe that Rho values for our pA site motifs are in general rather high, with their median rho value being 0.77, higher than the median value for promoter-associated motifs (0.71) or TF motifs (0.68). This likely reflects, in part, their location. 3' motifs are generally found within transcribed regions, which are systematically enriched for functional sites compared to untranscribed regions. Rho values for the pA site associated motifs are also often greater than for shuffled comparison motifs, with 11/24 pA motifs showing evidence for increased functional sites compared to shuffled motifs. Surprisingly, this is not true of the canonical CPSF motif, which may be because shuffled variants of 'AATAAA' within pA sites often overlap functional A-T rich motifs. Also surprising is the lack of functional enrichment in the 'unknown proximal' category of motifs, however this may result from the relatively low prevalence of most these motifs, which would make signatures of selection harder to detect. It may also reflect more tissue specific effects of pA site associated motifs – to the extent that factors like Tra2 or elav are specifically expressed, their motifs may occur, but be non functional, in RNAs with non-overlapping expression patterns – in effect resulting in false positive motif occurrences, which could also be caused by motif matches being frequently non functional for other reasons, for instance due to a requirement for other motifs.

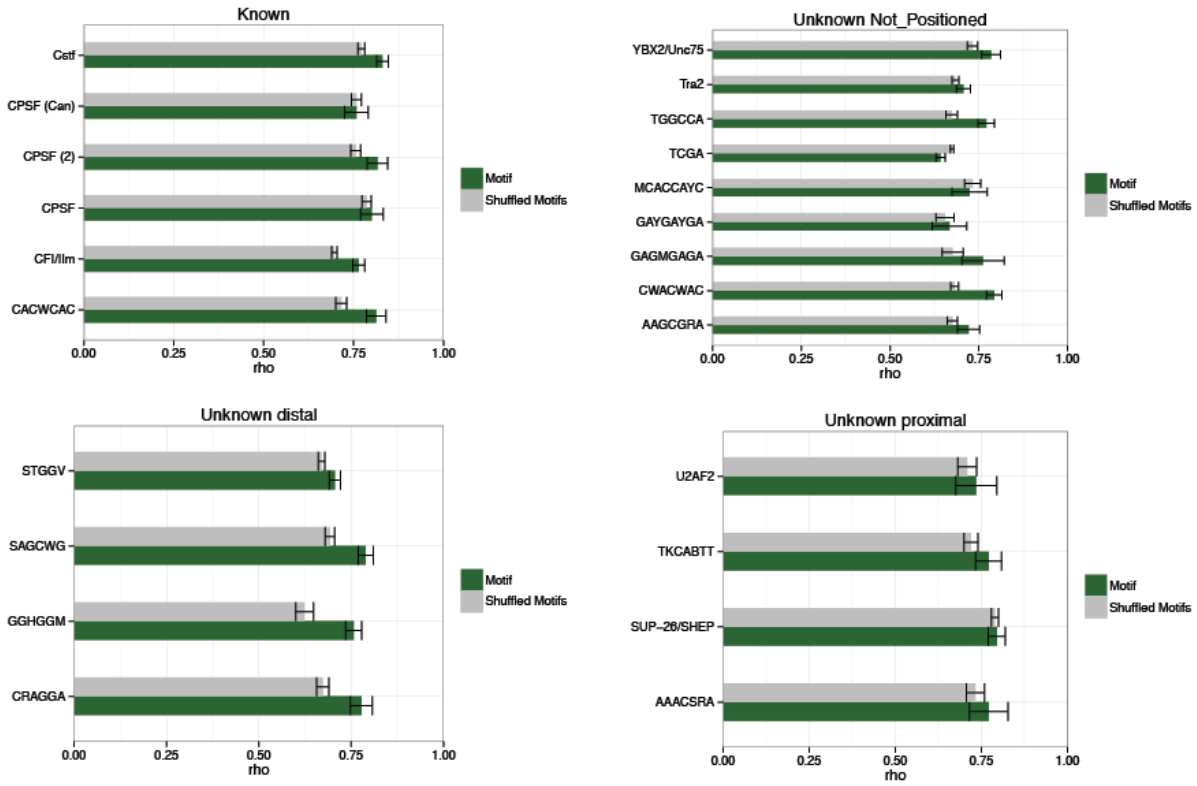


Figure 3.13 Rho values for pA site associated motifs

Proportion of functional sites (x-axis) for pA site association motifs (green) vs. their shuffled control versions (grey). *De novo* discovered motifs (using all pA sites) were grouped into 4 classes; known (previously characterised pA motifs), unknown not-positioned (relative to the pA cleavage site as per Centrimo), unknown distal (positioned further than 30bp from pA sites), and unknown proximal. Note that some motifs with well-characterized functionality, such as the canonical CPSF motif, are not more enriched than shuffled versions. This may reflect limitations of the shuffled motifs as a negative control, particularly for low complexity motifs, or the fact that since only motifs within Centrimo enriched regions were counted, shuffled motifs are likely to still overlap other functional motifs.

3.4 Discussion

In this chapter I have explored the sequence features associated with *Drosophila* transcription start sites, polyadenylation sites, and transcription factor occupied regions. One striking outcome of my work is the relatively poor performance of almost PWMs in the literature, including motifs identified by multiple, independent methods. Of the 1025 PWMs tested, just 78 passed my filters for AUC and enrichment score. This poor performance is relatively unsurprising - PWM's and consensus motifs alone are known to be poor predictors of transcription factor binding (e.g. Ernst et al 2010). However it is important to recognize that the process of de novo motif discovery is, even in theory, limited in its effectiveness. The models underlying *in silico* discovery tools typically fail to capture the properties of real sequence, but more problematically, the amount of information available in biological datasets is simply inadequate. Since only a finite number of binding sites will exist in the genome, and the space of possible PWMs is huge, accurately pinning down the properties of a TFs binding from ChIP data alone is inevitably error prone (Simcha *et al* 2012). This problem is unfortunately more pronounced in organisms like *Drosophila* with smaller genomes. Our quality assessment pipeline is more stringent than most tests of PWM accuracy in that it uses actual DHS sequence as control sequence, rather than randomly generated sequence. It therefore excludes many PWMs that are statistically enriched relative to more artificial control sequences, such as those generated by low order markov models.

By collecting and curating a set of high quality PWMs in *Drosophila* , my work will facilitate future studies of transcription factor binding and function in *Drosophila* . However the limited number of factors for which high quality motifs could be selected also emphasizes the limitations of PWM based models. PWMs are not the only way of summarizing transcription factor binding, and some alternative methods have shown significantly better performance (Alipanahi *et al* 2015) by using models of binding that do not assume a linear sum of contributions from each base pair. PWM models remain useful in part because their simplicity allows them to easily be

combined with other frameworks (e.g. Pique Regi *et al* 2010, Das *et al* 2016) and in part because of their computational tractability and interpretability. An obvious next step from my work would be an attempt to derive thresholds from less arbitrary criteria than 50% sensitivity. More recent efforts to use PWMs have sometimes gone beyond the use of PWM scores to derive scaled scores for transcription factor binding that more accurately reflect their biochemistry (Ma *et al* 2015). Another means of setting thresholds would be the joint modeling of sequence variation in conjunction with binding affinity – in principle conservation and allele frequencies should provide a means of dividing functional from non-functional motif matches.

The analysis of selection in transcription factor binding sites that I have performed here provides an interesting comparison to that performed in humans by Arbiza *et al* (2013). One striking difference is the frequency with which selection is observed. Aribiza *et al* found an average value of rho of 0.33 for their transcription factor binding sites. Of for my 78 high quality motifs, the average is 0.75. A truly accurate comparison would require a join analysis with equivalent thresholds for PWMs etc., however this large difference implies that more selection is detectable in the *D. melanogaster* genome that in the human genome. One simple explanation for this is the larger population size (Lynch 2007, Spivakov *et al* 2012). Another is that the *Drosophila* genome simply possesses less non-functional sequence by virtue of its regulatory elements being smaller and more compact. Distinguishing these two possibilities could be possible through the study of model organisms with large population sizes but larger genomes, such as sepsid flies (Peterson *et al* 2009).

Also striking is the presence of positive selection within many transcription factor motifs. These results expand previous findings for a small number of well characterized CRMs in *Drosophila* (He *et al* 2011) and for many in human (Arbiza *et al* 2014), indicating that transcription factor binding sites are under positive selection. The presence of positive selection within TFBS is significant since models of neutral evolution predict far less turn over in TFBS than is actually observed (Durret *et al* 2008, He *et al* 2011). One explanation for this higher turnover, which has been proposed, is that positive selection drives the evolution of new binding sites at rate higher than neutral evolution. This would predict that some existing

binding sites would show signs of the positive selection under which they evolved. My findings therefore support a model in which positive selection drives the gain and loss of binding sites in CRMs. This process may be particularly important in *Drosophila* due to the large population size, which strengthens the importance of selection relative to drift, and the fluctuating positive selection which the *D. melanogaster* genome appears to undergo (Mustonen 2007).

The analyses I have done here on selection within TFBS would benefit from many of the same refinements I have suggested above – further work to determine optional thresholds for PWM scores would give more accurate estimates of the degree of selection operating in *Drosophila* CRMs, and would allow more accurate comparison with equivalent human sequences. The correlation between motif quality and the presence of selection that I have observed underscores the importance of high quality PWM data for such studies.

The promoter-associated motifs that I have described here are promising candidates for future studies. It should be relatively easy to test each one for promoter activity using constructs similar to those used in enhancer assays. The evidence I have collected here on their shape preference and enrichment for functional sites should prove useful in guiding these experiments. I present more evidence in the subsequent chapter that these motifs are enriched for functional variants. The pA site associated motifs, similarly, present promising candidates for validation experiments, as well as similar caveats. Several of these were however present in sufficient numbers to allow the detection of their enrichment for functional variants (see chapter 4), which strengthens the case for their being *bona fida* functional motifs.

4 Mechanistic Analysis of eQTL in *Drosophila melanogaster*

Having carried out analyses of sequence features associated with promoters, transcription factor binding sites, and polyadenylation sites, we wished to use our identified sequences features as tools to understand the variation of gene expression in *D melanogaster*. Previous attempts to link sequence features with eQTL have been hampered by the large LD block size in humans (Reviewed in Gilad *et al* 2008, Gaffney *et al* 2012, Brown *et al* 2012) or by the lack of appropriate chip and sequence motif data in the model organism (Francesconi and Lehner 2014) By taking advantage of the small LD block size in *Drosophila* , we sought to gain insight into causal mechanisms behind eQTL, and assess the similarity of the mechanisms at work in *Drosophila* , as compared to human.

My analyses made use of several different classes of eQTL – called on both the CAGE and 3' Tag-seq expression data (see Fig 4.1). Both sets of QTLs were called using the Limix software package (Lippert *et al* 2014), however both QTL sets were analyzed in different ways. The CAGE QTL – or tssQTL – were analyzed with a focus on the distribution of tags within TSS. As such, the phenotypes used were not simple expression levels. Phenotypes were instead generated as follows: the CAGE signal was first divided into 1kb windows (see materials and methods). This resulted in an 81 x 2000 matrix for each window, with 81 being the number of lines tested, and 2000 being twice the number of basepairs in each window – i.e. one column per basepair/strand. Principal component analysis was then applied to each of these windows and the first 3 principle components were used as phenotypes, reducing the dimensionality of each line's phenotype from 2000 to 3. These 3 PCs were then used as phenotypes in a multi-trait linear model that accounted for background genetic variation as well as time point specific effects. The waveQTL algorithm (Shim et al 2015) was then used on the QTL identified as significant in this principle component analysis. WaveQTL decomposes the effects of QTL by projecting their

effects on to an alternative space of wavelets, which then allows Bayes Factors for spatially restricted effects, such as those effecting only one strand, to be detected. This strategy allowed the detection of traditional eQTL, but also the detection of QTL affecting the distribution of, rather than number of, tags at a TSS. CAGE QTL were therefore classified (using Bayes Factors from waveQTL) into Directional QTL, which correspond to traditional eQTL and affect the number of tags in a CAGE Window, Redistribution QTL, which affect only the distribution of tags, and Mixed – which affect both.

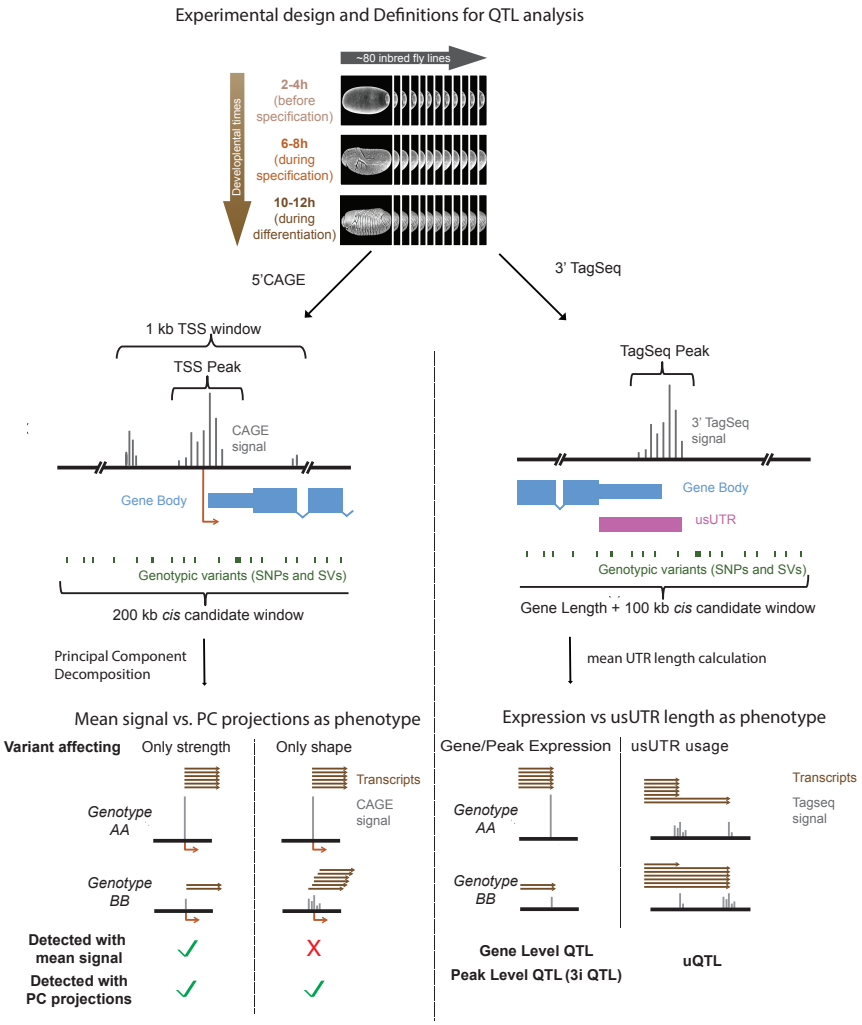


Figure 4.1 Various types of QTL used in the study.

Schematic explanation of the various types of QTL discussed in this chapter. Embryos from DGRP lines were collected at 3 time points (top) for both the 3' Tag-seq and CAGE assays. The CAGE data (left) results in reads from the 5' end of transcripts, so that it can be used to precisely measure TSS. The QTL strategy for tssQTL took advantage of this by decomposing the spatially distributed genetic variation at TSS into 3 principle components and using these at inputs for QTL calling. wavelet decomposition of these QTL was subsequently used to define sets of Redistribution and Directional tssQTL. The 3' Tag-seq data (right) results in reads from the 3' end of transcripts. We called QTL both by aggregating clusters of these reads (pA sites) on a gene by gene basis (gene 3' Tag-seq eQTL) and on individual pA sites (3i QTL). We also called QTL by weighting the usUTR length associated with each pA site by it's expression, and using this as a phenotype, resulting in uQTLs, which alter the gene's mean usUTR length.

In our 3' Tag-seq data, the subtypes of eQTL comprised eQTL – those called on the sum of a gene's pA sites, 3i QTL – those affecting individual peaks, (which may or may not significantly effect overall gene expression) and uQTL, those affecting a gene's mean usUTR length. The 3' Tag-seq QTL did not allow the same analysis of shape that the tssQTL allow (the lower resolution of the 3' Tag-seq protocol would make this impractical). They do however, in the case of uQTL, allow for the analysis of isoform usage, which the tssQTL, being called on individual windows and only later linked to genes, do not.

Thus, the QTL data sets in this study represent a survey not only of variants associated with gene expression variation, which are well studied in human, but also of the mechanisms underlying variation in transcription start site shape, and of 3' Isoform usage. Different classes of QTL show distinct mechanisms, and each provides a window into a different aspect of the genetic variation affecting the transcriptome.

To analyze the QTL, I have made use of datasets gathered and created in previous chapters, and demonstrate that the intersection of genomic datasets with eQTL data can be a powerful tool for understanding the mechanisms of gene regulatory variation. We have shown that many eQTL disrupt putative transcription factor, promoter, and pA site associated motifs. We have also discovered that eQTL in *Drosophila* regulatory regions are subject to epistatic interactions, and that this epistasis can act both at the level of gene expression, and cell to cell variability. We have also discovered and that broad promoters are subject to additional functional variation affecting the distribution of TSS over their breadth. These results represent a substantial contribution to our understanding of gene regulatory variation.

4.1 Assessing QTL enrichment within genomic features using logistic regression

Analyzing the enrichment of a given feature for eQTL is non trivial, because of the many confounding factors which influence the likelihood of a variant being assigned QTL status within a given genomic region. The first and most obvious of these confounders is the polymorphism rate within a feature – the more variants

that are present with a feature, the more QTL can be found in a feature. This will obviously vary by the size of the region, but also by mutation rate and selective constraints. Less obviously, the minor allele frequency (MAF) of variants within a feature will also influence the likelihood of their being marked as QTL – the analysis will have less power to detect functional mutations at a given effect size when their frequency is low. Similarly, the expression level of the gene being tested for QTL will also determine statistical power (see Fig 4.2). Finally, proximity to the point of measurement (i.e. TSS or 3' Tag-seq Peak), is will often obscure other signals unless controlled for, where features differ in their average proximity to the TSS.

To deal with these issues and estimate QTL enrichment I elected to use a modified version of the logistic framework used by Brown $et\ al$ (2012). The slope of this logistic regression can be thought of as giving a logs odds ratio (which I present in odds form) for the probability of a given variant being a QTL, given that it is or is not within the genomic feature. I use this framework (see materials and methods) to derive the 'enrichments' discussed in the following sections.

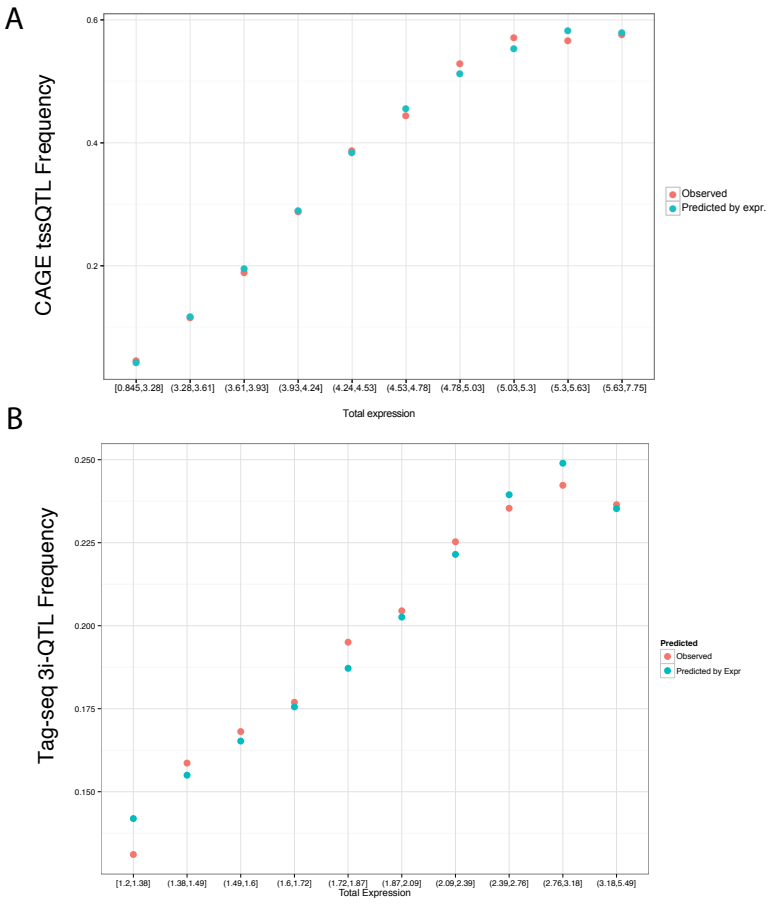


Figure 4.2: Expression Levels are a strong predictor of QTL frequency for both 3' Tag-seq and CAGE QTL.

Plots showing trend for higher QTL frequency with increasing expression. Genes were binned into deciles by expression (x-axis). The y-axis shows proportion of QTL per bin, red dots show observed frequencies of QTL per bin for tssQTL (A) and Tag-seq QTL (B). The blue dots show the predicted average per bin using a simple logistic regression model with expression as a covariate.

4.2 Genomic distribution of eQTL

Previous studies in human have shown that eQTL are in general enriched within transcribed regions (e.g. Kwan *et al* 2008, Pickrell *et al* 2010), close to transcription start sites, and in exons relative to introns, (Veyrieras et al 2008). Enrichment with transcribed regions is relatively unsurprising – many enhancers are located

intronically, and variants in transcribed regions carry the possibility of affecting post-transcriptional regulation as well. Similarly, the TSS is an obvious focal point for variants affecting gene expression, given that it is the focal point for transcription initiation.

Mapping the locations of our tssQTL relative to their target gene (Fig 4.3) showed that our QTL, like those in human were also clustered around TSS.

3’ Tag-seq 3i QTL (Fig 4.3), in addition to clustering around TSS, also showed clustering around transcription termination sites, in agreement with their potential involvement in polyadenylation and cleavage. This was particularly true of uQTL, suggesting their involvement in 3’ UTR biology. The trend was less pronounced for eQTL, though this was likely in part a result of filtering these for variants overlapping pA sites.

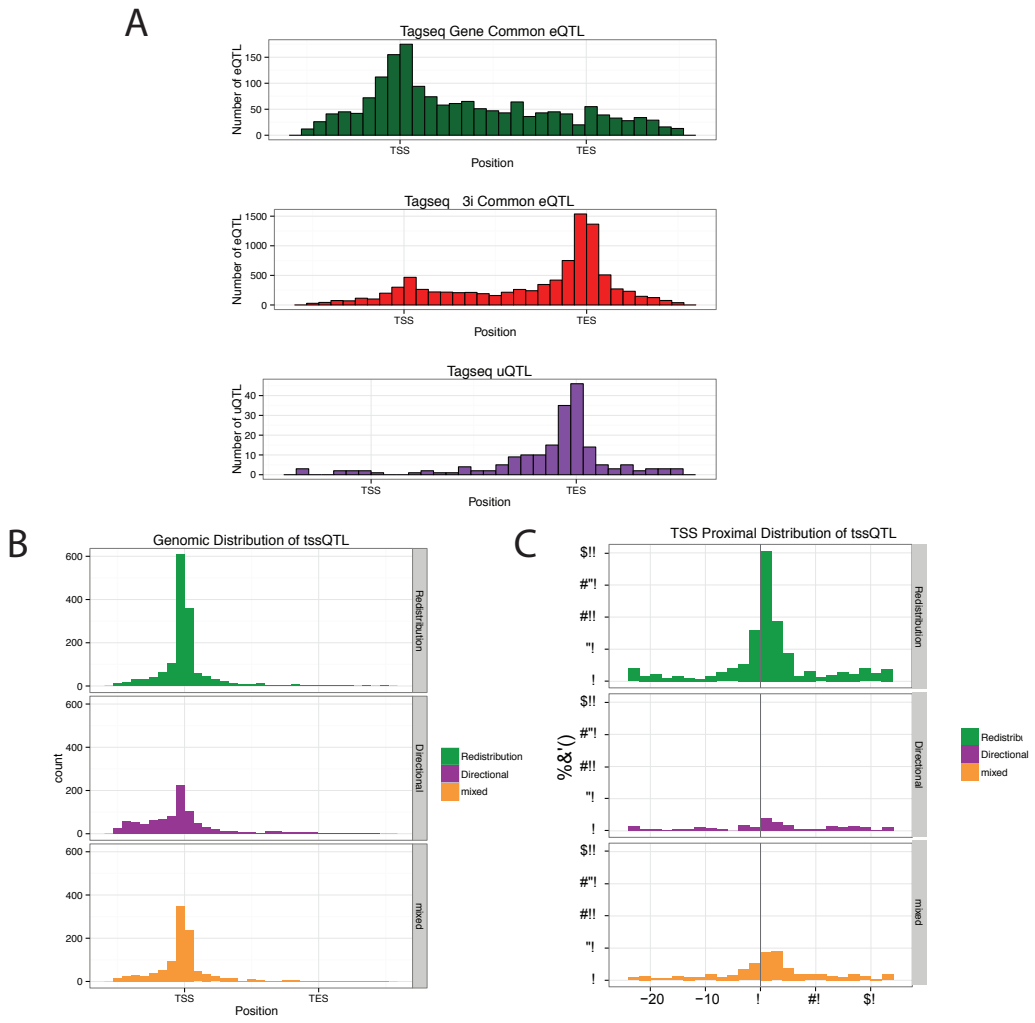


Figure 4.3: Positional distribution of QTL.

A) Distribution of 3' Tag-seq QTL over the gene body. Metagene profiles were constructed using Flybase gene annotations, with the gene's largest transcript used for each QTL. B) Same plot for tssQTL. C) Same plot for TSS QTL, zoomed in on +/- 25 bp around the TSS, where the TSS is defined as the CAGE-tagged site most affected by the tssQTL. Note the greater clustering for Redistribution QTL. tssQTL are enriched around promoters and associated genomic features. 3i QTL are enriched around TSS, but also around TES.

Interestingly, many stage specific Gene QTL are located outside of TSS. Of the 314 for which one variant was at least an order of magnitude higher than others, 258 are outside TSS, 90 overlap a region bound by two or more TFs, and/or a DHS. We were also surprised to observe that 1/3 of all gene-eQTL were located > 10kb from

the affected gene's TSS, with 28 greater than 50kb away, and the furthest 81kb. These results points to a greater role for distal regulation in *Drosophila* than previously identified, in agreement with orthogonal data from two recent studies (Kvon *et al* 2014, Ghavi-Helm *et al* 2014). Furthermore, by providing examples of specific distal regulatory links, they provide a useful set of targets for further experimental investigation.

TssQTL are in general more clustered around TSS than the 3' Tag-seq QTL, with 85% of Redistribution tssQTL, and 63% of Directional tssQTL, within 1kb of a TSS. We observe that even within this zone, Redistribution QTL are more highly clustered around TSS (Fig 4.3) than are Directional CAGE QTL, with Mixed QTL showing an intermediate phenotype. This supports the idea that Redistribution QTL represent variants affecting the biology of transcription initiation itself, whereas Directional QTL represent all variants which result in different steady state transcript levels – hence potentially also including variants in distal regulatory regions, RNA binding protein motifs etc.

Moving to a logistic regression based measure of enrichment (Fig 4.4) we confirm that QTL are significantly enriched within TSS, with the notable exception of Directional CAGE QTL, again emphasizing the distinct biological mechanisms that underlie the different sets of CAGE QTL. This is true of both Gene and 3i QTL. We also observe that similarly to human eQTL, our *Drosophila* eQTL are enriched in introns relative to coding exons, with Redistribution and Mixed CAGE QTL showing lower enrichment in both. 5' UTRs show in general similar patterns of enrichment to TSS, however 3' UTRs show markedly different enrichment levels between QTL types. Of the relatively few CAGE QTL in 3' UTRs, only Directional QTL are enriched, which is no doubt a result of the fact that change in 3' UTR sequence tend to affect stability of transcripts and hence overall levels, rather than transcript initiation sites. 3' Tag-seq QTL in general show enrichment in 3' UTR, with 3i QTL showing more enrichment than Gene QTL, and QTL with stage specific effects showing less enrichment than QTL with effects common to all developmental stages. This observation immediately suggests that the mechanisms at work in stage specific and common QTL are to some extent distinct. Intuitively, one would expect changes to 3' UTR sequence to show less developmental specificity than effects originating in

distal regulatory regions since enhancers often show spatiotemporal specificity, an expectation that our data support.

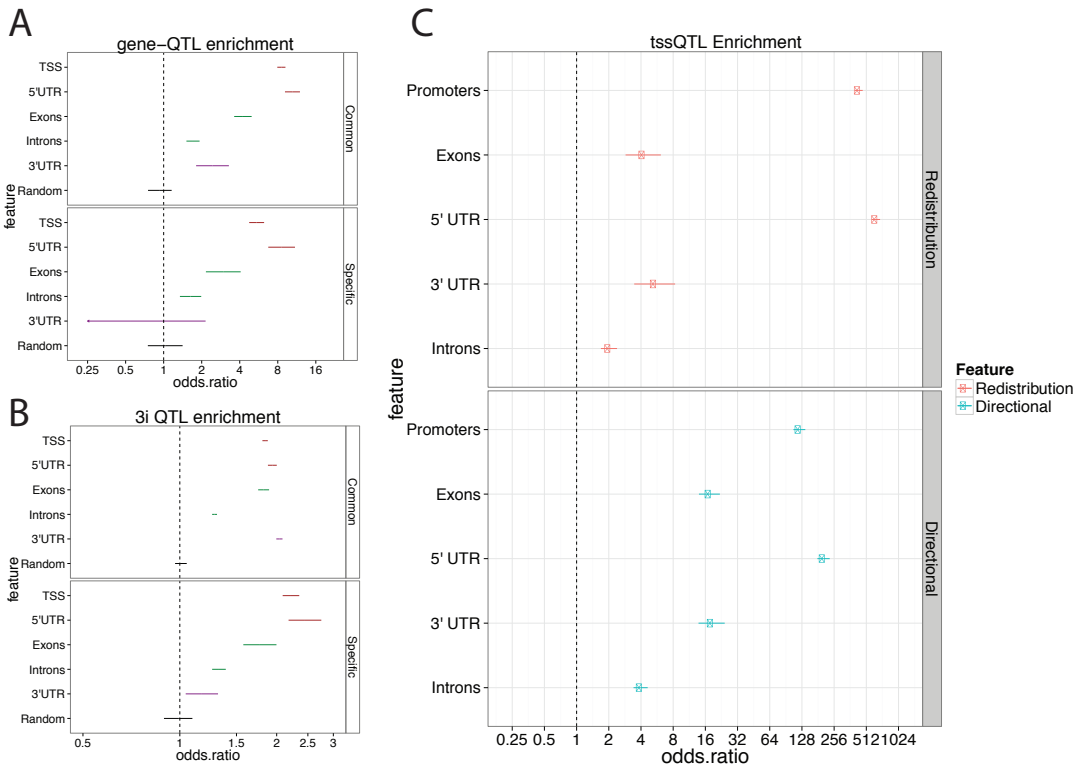


Figure 4.4: Enrichment of 3' Tags-eq QTL, 3' QTL and tssQTL in various gene regions.

A) 3' Tag-seq gene-eQTL and their enrichment (x-axis) within gene-annotation features, relative to all tested SNPs. As expected, TSS and 5' UTRs are enriched for QTL, with exons slightly enriched over introns. "Random" denotes enrichment of randomly chosen negative regions, included as a negative control. B) Same but for 3i QTL. Enrichments for 3i QTL are very similar, although note that time point nonspecific 3i QTL are enriched within 3' UTR as well as 5' UTR. This effect is absent for 3i QTL, which may reflect different biological causes of the two sets, or a greater proportion of the time point common 3i QTL being artifacts of mapping bias. C) Same for tssQTL. Enrichments are again as expected, but stronger, reflecting the tssQTL's greater tendency to cluster near TSS. Bars represent 95% profile confidence interval of odds ratios, calculated using logistic regression. See materials and methods for details of model.

4.3 eQTL presence within *Cis* regulatory module features

A number of papers have observed that genomic features associated with CRMs are enriched for the presence of eQTL. For instance, Gaffney *et al* (2012) find that features like DNase I hypersensitivity hotspots, and various histone marks, show 2-4 fold enrichment in causal SNPs, and that these features become more heavily predictive outside of TSS proximal regions. We used our logistic regression framework to assess QTL enrichment in features associated with QTL. These included DNase hypersensitive regions, regions occupied by Transcription factors (see materials and methods), and transcription factor motifs bound by occupied by the relevant factor. We divided our QTL into a distal set – those > 1kb from a TSS, and a proximal set, those within 1kb of a TSS. 35% of these distal genes QTL overlap a known DNase hypersensitive site, a peak of two or more transcription factors, or a peak of H3K27ac or H3K4me1, while not overlapping a peak of H3K4 tri-methylation (which would indicate unannotated promoters), suggesting that these QTL may act by affecting enhancers.

In agreement with human studies, we find that these features, which are (imperfect) markers of regulatory sequence, show enrichment for 3' Tag-seq eQTL, with DNase hypersensitive regions showing the strongest enrichment. Transcription factor bound regions also show enrichment for QTL. Puzzlingly, transcription factor motifs do not show significantly higher enrichment than chip bound regions, of which they are a subset. This may in part be due to the necessarily imperfect nature of binding site calls based on simple PWMs.

Paradoxically however, it could also reflect the fact that the TFBS regions may be *more* functional than the ChIP bound regions. Indeed, if poor PWM calls were the sole reason for the lack of QTL enrichment in TFBS, we would not expect to see the signs of increased natural selection we observe in these sights. If mutations within these sites are prone to cause effects with a large impact on the CRM's regulatory potential, they may be less likely to reach a high minor allele frequency, and hence would be underrepresented in our QTL data – which cannot detect the effects of low frequency mutations.

The same analyses carried out on the CAGE QTL yielded similar results for the proximal set of features (Fig 4.5). However we observed no significant enrichment, either for Redistribution or Directional QTL, for the distal set of CRM features. In the case of Redistribution QTL, this likely again reflects mechanisms that center on the TSS and the initiation machinery itself. In the case of Directional QTL, this is more puzzling, but given that the confidence intervals still allow for modest enrichment in CRM features, may simply reflect the low number of distal QTL in this set.

We also identified 43 Gene QTL that overlap a set of enhancers with experimentally validated regulatory activity. This set of enhancers did not show significant enrichment for QTLs on a global level, which may be a result of the regions' large size (median size = 2kb). This large average size is a results of the enhancers being derived from studies carrying broad scale studies of regulatory regions, rather than focused analysis of sequence motifs.

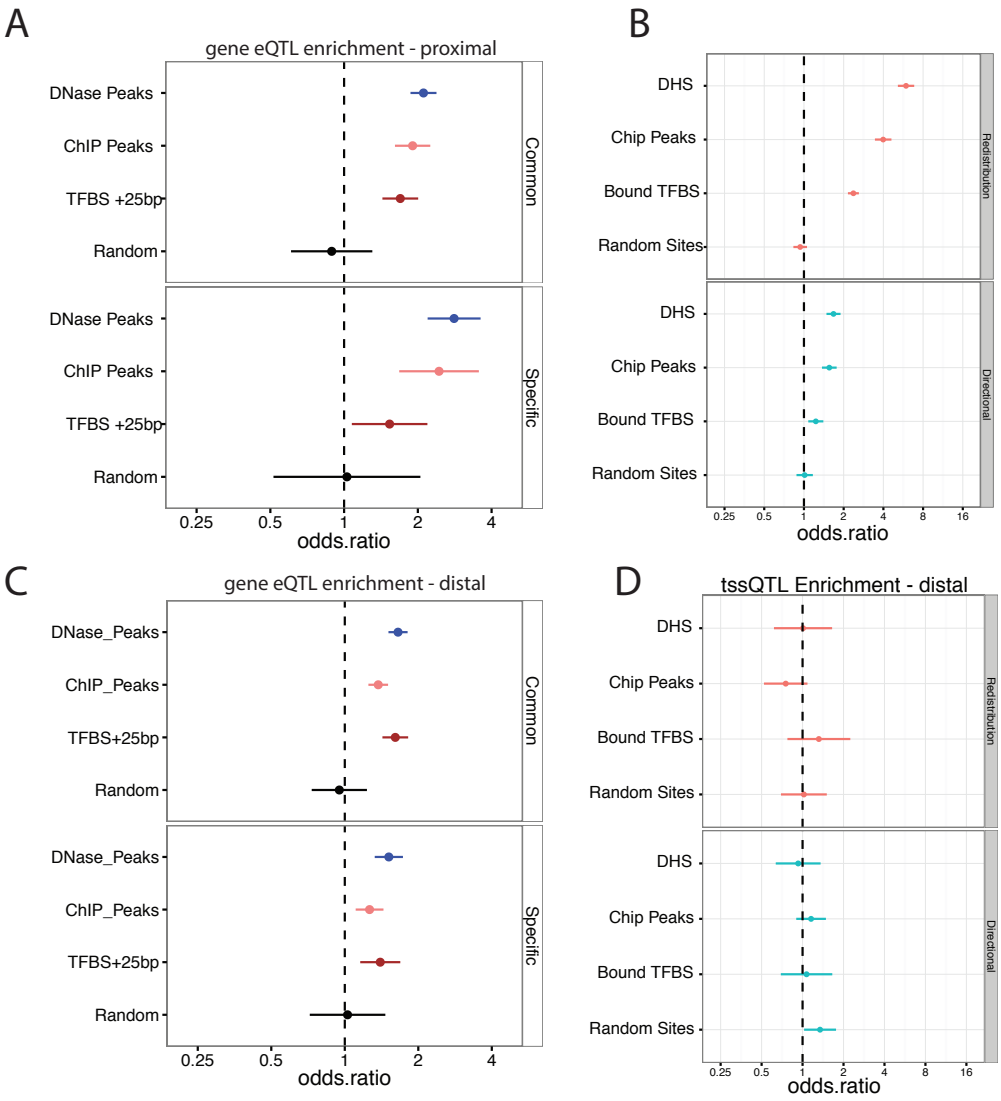


Figure 4.5 Enrichment of 3' Tag-seq QTL and tssQTL in CRM related features.

A) 3' Tag-seq QTL and their enrichment (x-axis) in DHS, ChIP peaks (the High Quality set defined in chapter 3) and the 25bp area around TFBS, (matches to the high quality motifs described in chapter 3). Enrichments shown are for distal (B,C) and proximal (A,D) regions, where a proximal variant is one within 1kb of a Flybase TSS for the tested gene. Bars represent 95% profile confidence interval of odds ratios, calculated using logistic regression. See materials and methods for details of model.

4.4 eQTL changing motifs

4.4.1 3' Tag-seq eQTL alter transcription factor motifs

The large collection of eQTL identified by our study includes, surprisingly, eQTL affecting major developmental factors (49 transcription factors have detectable eQTL), as well as their motifs. We assessed changes to transcription factor binding sites using a custom pipeline. The pipeline works by constructing local haplotype sequences, scanning both for the presence of motif matches using our set of 50 high quality PWMs, and assigning the best score in each haplotype to the corresponding variants. This process represents, to our knowledge, the most efficient means of assessing PWM changes directly from Variant Call Format data without the requirement for personal genome construction, and is capable of dealing with indels, and combinations of variants. Previous studies have examined QTL affecting sequence motifs and documented a correlation between the strength of PWM change and the likelihood of a variant being functional. For instance, Ding *et al* identified motifs associated with changes in CTCF occupancy, and noted that variants in basepairs of the CTCF motif with high information content were more likely to be QTL (Ding *et al* 2014). Biochemical studies of the transcription factor hunchback (He *et al* 2011) also indicate that a score change of -1 or more to its PWM tends to result in alterations in biochemical affinity. In order to select variants strongly affecting PWMs therefore, we counted only changes of three or more points to the PWM score between haplotypes. This threshold, while arbitrary, did not strongly affect our results when varied.

Examining the variant with the lowest p-value for each QTL used this pipeline to detect changes in transcription factor motifs. Our analysis showed that many of our 3' Tag-seq eQTL disrupt binding sites for known developmental transcription factors such as Snail (Fig 4.6a), a transcriptional regulator involved in early mesoderm patterning in the *Drosophila* embryo (Rembold *et al* 2014), and twist, a transcriptional activator which also functions in mesoderm development (Sandmann

et al 2007). In all, after filtering by a score change threshold and occupancy of the relevant factor (or DNase sensitivity in 'created' motifs present in the alternative but not reference genotype) we observed 95 instances of destroyed transcription factor motifs (three of which we selected for experimental analysis below), and 183 instances of created transcription factor motifs (relative to the reference sequence) (Fig 4.6b). The asymmetry between the numbers of created and destroyed motifs is in part a result of the different means used to filter them – since ChIP datasets are collected on the reference genotype, we could not apply the same filters to created motifs, and instead used a less stringent criterion that they overlap DNase sensitivity peaks. Notably, the number of motif destructions in DNase peaks – 157 – was similar to the number of motif creations.

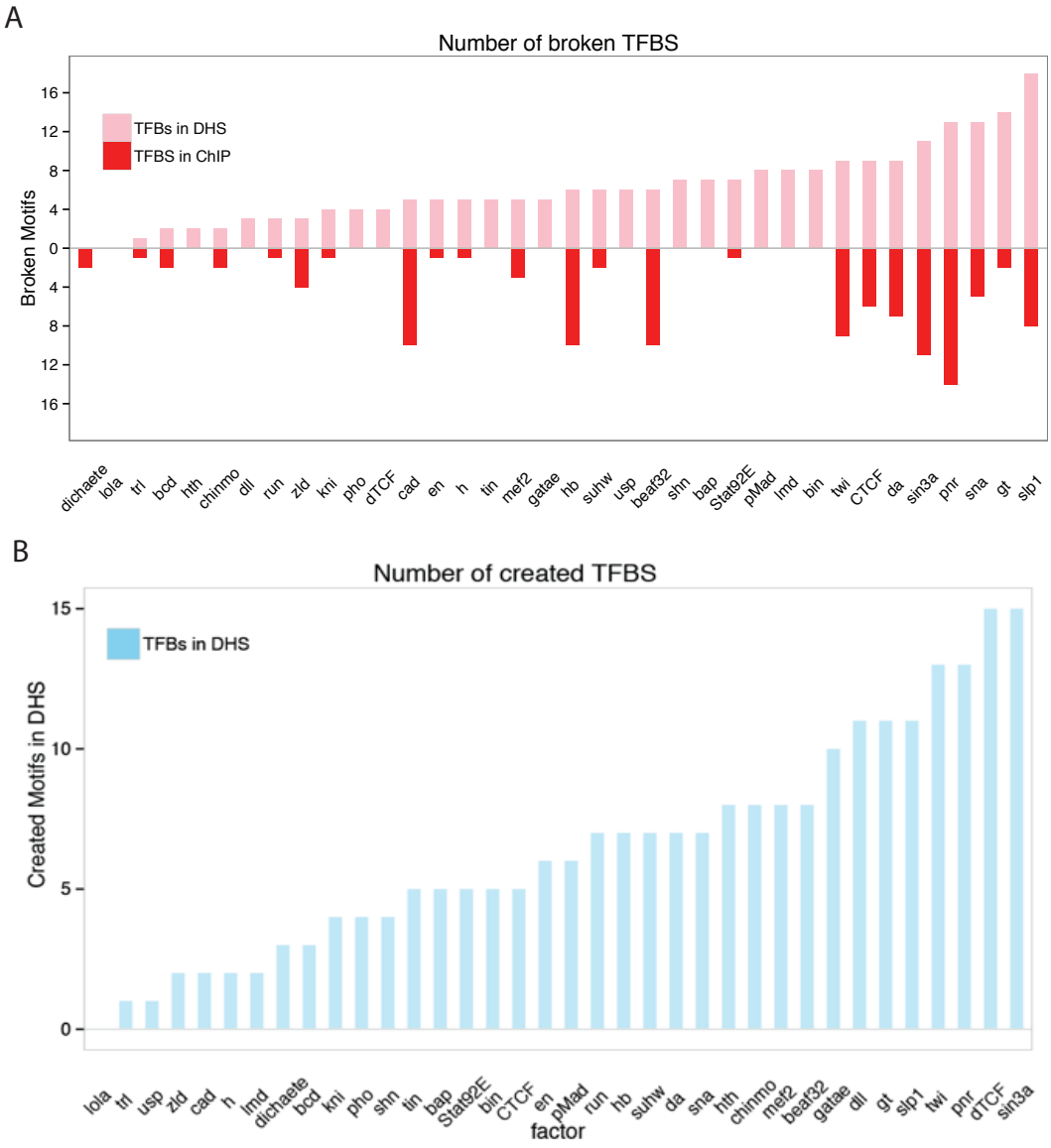


Figure 4.6: 3' Tag-seq eQTL destroy and create motifs

Numbers of TFBS that are broken (A, red) or created (B, blue) in TF bound regions (dark red) or DHS (light red/bue), by 3' Tag-seq Gene QTL. For each QTL, the variant(s) with the strongest p-value were chosen. Alternative and reference sequences for these were then scanned using Patser. Cases in which the score for this motif changed by 3 points or more, and either alternative or reference had a significant motif, were counted. Since ChIP data likely reflects occupancy on the reference sequence, we used DHS data to filter instances of motifs being created, assuming that functional motif creations would mostly occur in regions with pre-existing regulatory activity. Only quality filtered PWMs (see chapter 3) were used.

Our results allowed us to attribute a significant but relatively small proportion of QTL (278/3767) can be attributed to recognizable changes in known transcription factor motifs. This figure would no doubt increase were we to increase the number of PWMs in our high quality set, or decrease the threshold at which motifs were called, however both of these steps would require lowering the quality of our motif calls and losing accuracy. The results illustrate the importance of genomic data quality in the analysis of regulatory variation, and also the fact that regulatory variation involves a wide range of mechanisms, such that even if the set of QTL is large, the set of variants attributable to any specific mechanism is still likely to be small. Nevertheless, our results provide a set of high confidence changes to TF motifs that can serve as a basis for experimental analysis of naturally occurring regulatory variation in *Drosophila* .

We also reasoned that if change in transcription factor affinity is the causal mechanism behind our QTL, we should observe correlations between the effect size of our QTL, and the magnitude of change to the PWM. We observed no such correlation for any of our PWMs, nor did we observe a significant bias towards positive/negative effect size for individual PWMs. This may reflect the relatively poor performance of PWMs, or simply a lack of power due to low numbers of motif changes for individual factors.

4.4.2 3i QTL alter pA site associated motifs

Having examined transcription factor binding in the Gene 3' Tag-seq eQTL, we then wished to examine the mechanisms at work in the 7124 common and 1241 stage specific 3i QTL. Because these QTL represent variations at the level of individual 3' isoforms, we reasoned they would show differing mechanisms (although it should be noted that some 3i QTL are also eQTL). Unlike eQTL, 3i QTL were not filtered for proximity to pA sites, (since so many of them are found proximal to peaks). We reasoned that omitting this filter, although it would increase the false discovery rate due to mapping bias, should do so in a manner independent of sequence motifs, allowing broad conclusions to remain valid. Exceptions to this would occur where

motifs were differentially present at the point of measurement – such as our pA sites - which is indeed the case for some motifs. Therefore, when calculating enrichments for 3i QTL, we included an additional binary control variable in the logistic regression, controlling for presence within a pA site. This control variable should prevent spurious enrichments due to mapping bias, but likely controls away some actual biological signal as well.

We first asked whether variants affecting the 24 pA-site associated motifs we discovered by de novo motif analysis result in 3i QTL. We found many instances of 3i QTL that either destroy or create one of these motifs Fig 4.7a,b. Particularly common are variants which affect CPSF binding motifs, which by disrupting the known cleavage machinery present an obvious mechanism by which 3i isoform usage could be affected. We next examined the global enrichment for 3i QTL of variants creating and destroying these motifs as well as the motifs of some known RBP proteins (Ray *et al* 2013) see Figure 4.8. Variants affecting canonical CPSF motifs are particularly enriched for 3i QTL, indicating that the many 3i QTL affecting them are not simply a result of their high prevalence in pA sites. Many of the novel motifs we identified as enriched near pA sites are also enriched for 3i QTL, including two motifs (AAACSRA and TKCABTT) that are proximally positioned around the point of cleavage. Interestingly, variants both creating and destroying these motifs were enriched for QTL, and to differing degrees. For instance, variants destroying Elav motifs are highly enriched for QTL, while those creating Elav motifs are not. This suggests that the creation of an elav motif alone tends to have little effect, perhaps because elav requires other sequence features, or other proteins, to affect gene expression. It could also be that the restricted expression pattern of elav permits variants that create Elav motifs to exist in genes that are not expressed within the nervous system (we tested an association between the expression pattern of elav motifs and the likelihood of their causing an eQTL, however no relationship was clear, which may simply be due to a lack of power – our test had a power of only 0.05 to detect the observed difference as significant).

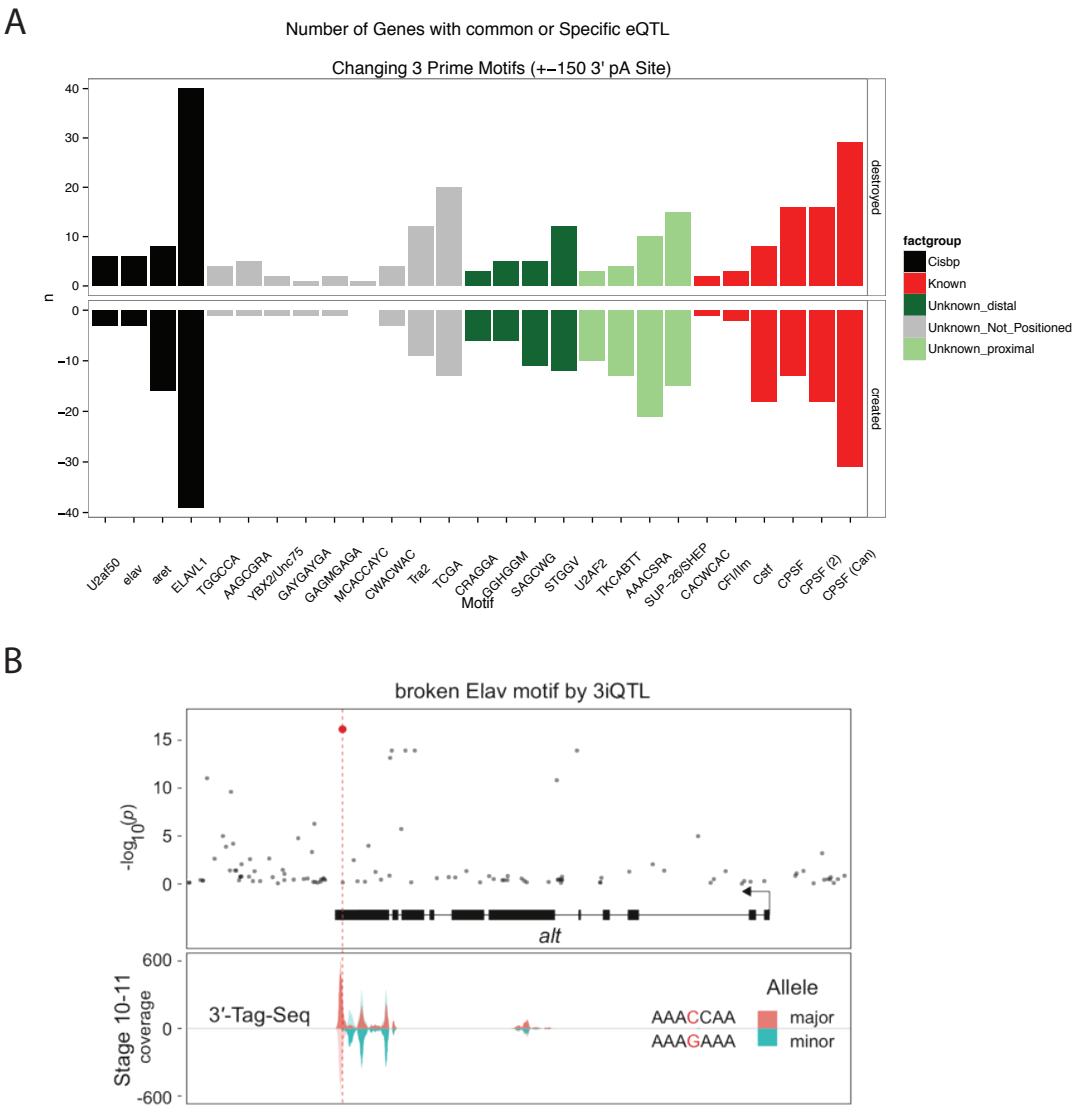


Figure 4.7: Motifs created or destroyed by 3i QTL.

A) histogram showing number of 3i-QTL either breaking (above x axis) or creating (below) motifs for known polyadenylation cleavage motifs (red) or *de novo* discovered motifs. B) Upper panel, Manhattan plot showing unadjusted –log10(P-value) for all tested variants within the extended *alt* locus. Red dot indicates the most strongly associated genetic variant which destroys an Elav motif. Lower panel, 3'-Tag-Seq data for the major (dark red, above) and minor (blue, below) genotypes (median coverage). For ease of comparison, the major genotype signal is shown in light red below and the minor genotype signal in light blue above.

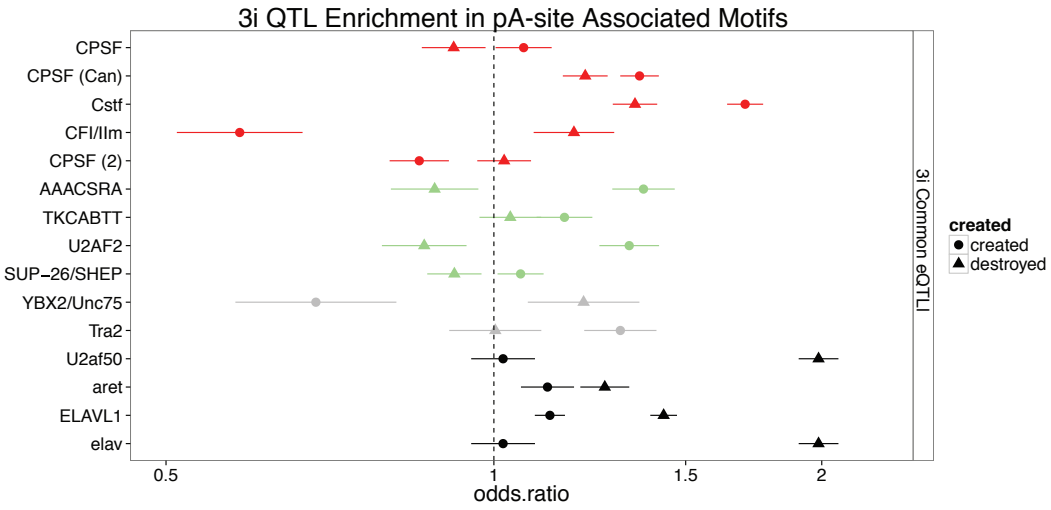


Figure 4.8: Motifs created or destroyed by 3i QTL.

Global enrichment of 3iQTL in pA site associated motifs assessed using multivariate logistic regression. Bars represent 95% profile confidence interval of odds ratios. See materials and methods for details of model. “Created” Motifs are present only the alternative genotype, while “destroyed“ motifs are present only in the reference.

By grouping our pA site associated motifs, we also observed differences in the effect sizes of eQTL affect motifs depending on whether a motif was created or destroyed. For example, motifs that destroy canonical PAS motifs (Fig 4.9c) are significantly more likely to reduce gene expression. This is likely because the resulting transcripts include genomic sequence not adapted for transcription, and thus less stable than slightly truncated transcripts that result from early PAS motifs. A trend in effect size is also detectable in variants affect binding sites for Tra2 (Fig 4.9a,b) - a well characterized splicing factor with roles in sex determination in *Drosophila* . A number of 3iQTL affect Tra2 motifs near splice junctions, but even more are enriched within the 3' UTR itself. These pA proximal Tra2 motifs have a stronger effect when disrupted, suggesting that Tra2 motifs might play a previously uncharacterized role in Poly-A site related RNA processing.

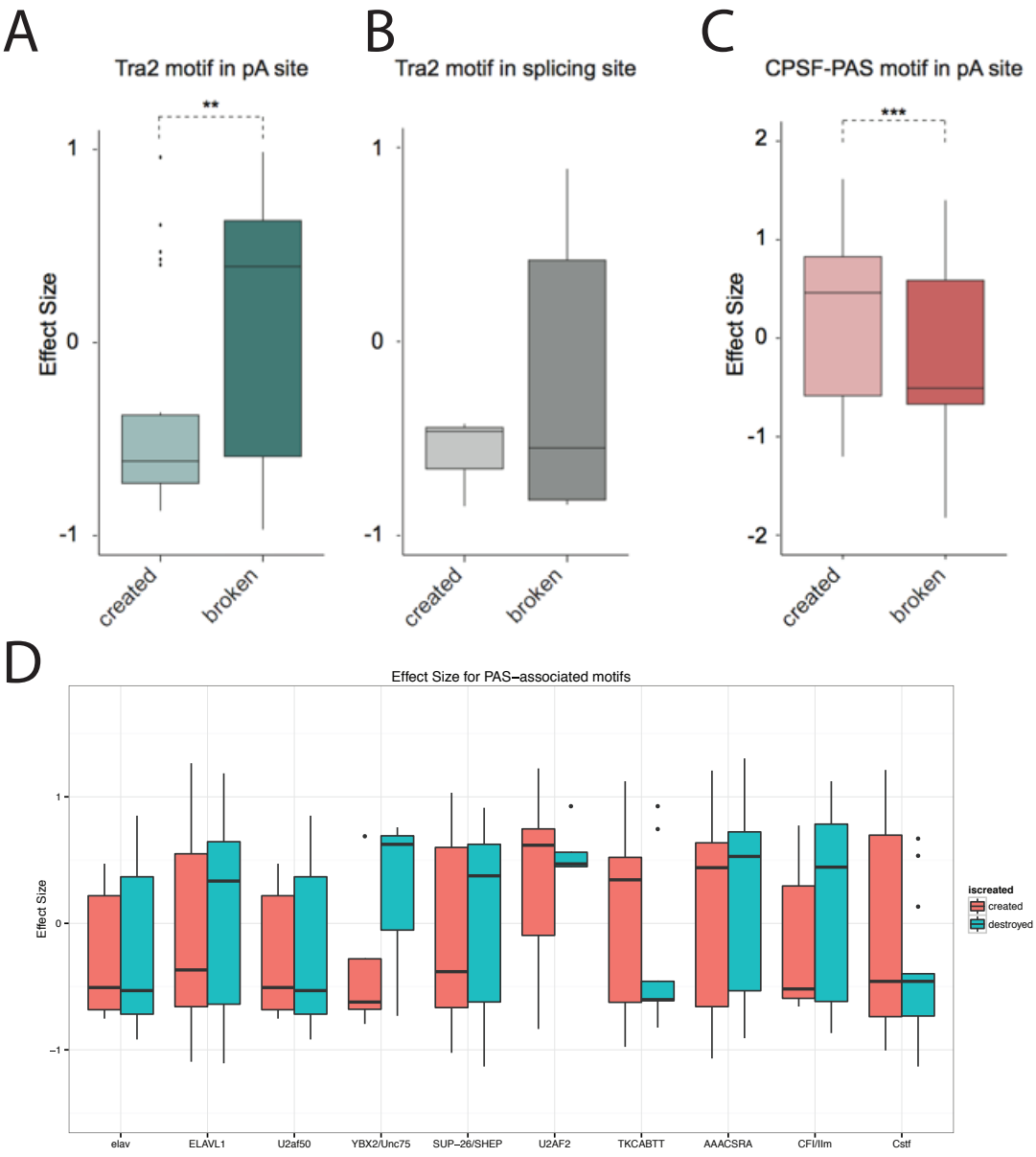


Figure 4.9 Relationship between effect size and motif destruction/creation.

Effect size (y-axis) for 3i QTL that break (motif absent in minor allele) or create (motif present only in minor allele) pA site associated motifs. 3i QTL that A) disrupt Tra2 sites near pA sites, show significant difference in effect size between created vs. destroyed motifs, while those near splice sites (B) do not. 3i QTL destroying canonical PAS motifs (C) also show a significant difference. D) Other motifs do not show significant differences after correcting for multiple testing ($p<0.05$ Wilcoxon signed rank test).

4.4.3 QTL altering length affect pA site associated, and RBP motifs

Elav is an RBP expressed in neurons, which inhibits 3' polyadenylation (Dai *et al* 2012), resulting in increased 3' UTR length of neuronal genes (Hilgers *et al* 2012). We observed that 3iQTL disrupting Elav motifs are often located close to the transcript end (e.g. Fig. 4.7b) and alter 3' UTR length. In the alt locus, for example, loss of an Elav motif results in a shorter 3' UTR in the minor allele, compared to the major, in keeping with its characterized function (Fig. 4.7b). We note however, that the direction of change in UTR length was not consistent between motifs creating/destroying elav motifs. This suggests that there are likely other proteins regulating 3' UTR length, and also that some 3i QTL may simply happen to overlap nonfunctional elav motifs.

We wished to systematically determine which variants were causing changes in UTR length, independent of overall peak expression. Further QTL analysis was carried out, in a manner similar to the eQTL, but this time using overall gene expression as a control variable, and the genes' mean unspliced UTR length values (see chapter 2) as the phenotype. In all, 764 genes had UTR-QTL, illustrating that UTR length varies substantially. Focusing on the variants that caused substantial changes in length (greater than 25bp) we identified 311 UTR-QTL, causing a mean length change of 57 bases. The QTL are equally divided between those that increase and decrease UTR length, and are distributed close to transcript ends. 78% of genes with UTR QTL also have 3i QTL, and almost half of these are associated with a different variant, illustrating the complexity of genetic influences on different phenotypes.

Almost 30% of UTR-QTL (uQTL) disrupt a motif resembling one bound by an RBP (Fig 4.10c). This includes three UTR-QTL that alter Tra2 motifs and eleven cases that alter Elav motifs. The single most common motif disrupted by UTR-QTL is the PAS motif, and 23 (7%) of UTR QTL affet canonical PAS or variant PAS motifs. YL-1 provides a good example of a UTR-QTL disrupting a PAS motif. The canonical PAS site disrupted by the single lead SNP in the minor allele leads to an increase in the usage of a distal polyA site, and hence to an increase in UTR length in lines with that SNP.

In all, these results suggest that, as with eQTL disrupting TF motifs, the mechanisms at work in UTR-QTL are diverse, such that even with many QTL, the number of variants affecting any given motif is small. As such, we lacked power to detect any global enrichments or effect size trends with UTR-QTL.

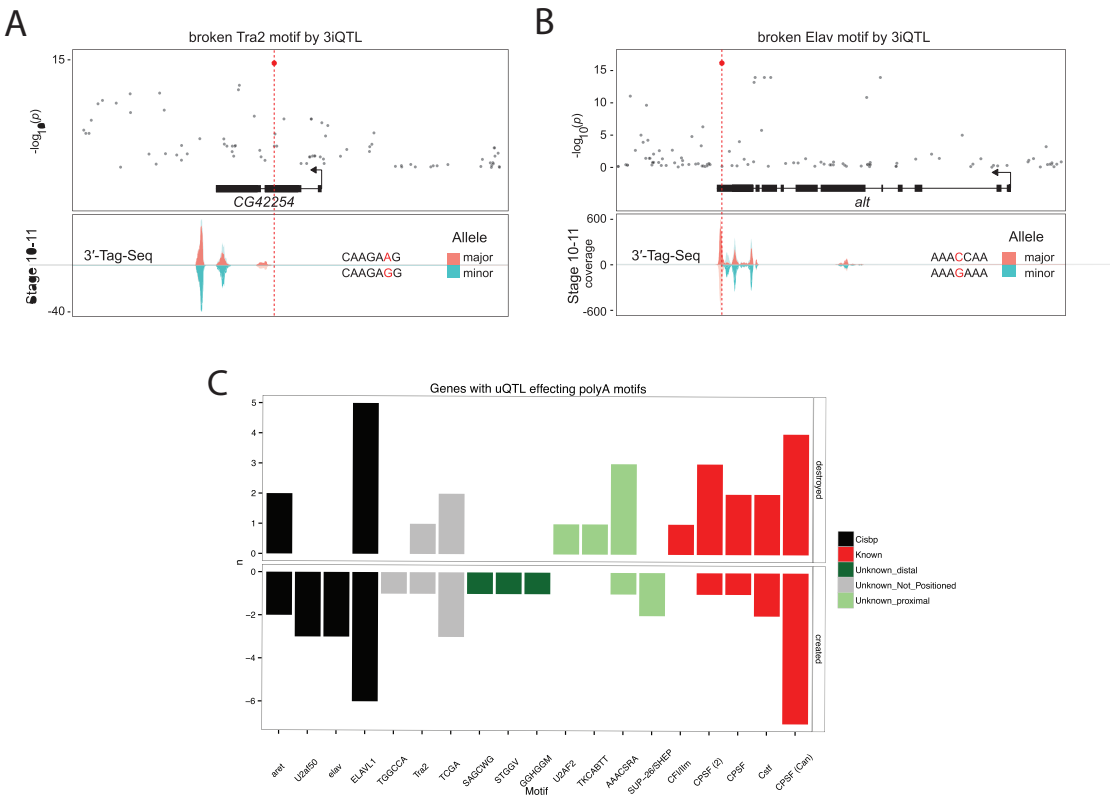


Figure 4.10: Motifs created or destroyed by utrQTL

A) Upper panel, Manhattan plot showing unadjusted –log10(P-value) for all tested variants within the extended *CG42254* locus. Red dot indicates the most strongly associated genetic variant, which disrupts a Tra2 motif. Lower panel, 3' Tag-seq data for the major (dark red, above) and minor (blue, below) genotypes (median coverage). For ease of comparison, the major genotype signal is shown in light red below and the minor genotype signal in light blue above. B) Manhattan plot for the extended YL-1 locus, which in which a uQTL disrupts a PAS (CPSF) motif. C) Number (y-axis) of uQTL either breaking (upper panel) or creating (lower) motifs for known polyadenylation cleavage motifs (red), *de novo* discovered

4.4.4 tssQTL disrupt promoter associated motifs

We reasoned that tssQTL could function by affecting binding motifs for promoter associated factors, and core transcriptional machinery. Individual promoter-associated motifs, since they are generally present in low numbers and at a fraction of genes, will not have enough QTL affecting them for global statements about enrichment or effect size to be made. We found that other sets of promoter-associated motifs (Ohler *et al* 2002, Fitzgerald *et al* 2006, Tiffin paper) also present

relatively low numbers of creation/destruction events and thus did not show any global enrichment trends. We therefore elected to group our discovered motifs together (a strategy similar to that used by Gaffney *et al* (2006). Reasoning that position with respect to the TSS could allow us to group motifs with similar functions (with e.g. the INR and Motif 1 motifs being grouped together). We also compared these sets to motifs without any preference in localization, strand, or promoter type. We also group variants as either creating or destroying motifs, or combine those groups into the 'changed' category. Fig 4.11 shows the resulting enrichment scores for the motif classes. We find that the upstream, downstream, and TSS classes are all enriched for QTL. Redistribution and mixed tssQTL are more highly enriched at TSS and downstream motifs, while Directional QTLs are only slightly enriched at TSS-positioned motifs, and clearly enriched at downstream motifs, for both creations and destructions. This suggests that the Redistribution QTL are more likely to directly affect motifs at the start site, which is in line with their mechanisms directly affecting the process of transcription initiation. The generally lower enrichment seen in upstream motifs may indicate that these motifs are less likely to be functional. The difference between upstream and downstream motifs could also be a consequence of the fact that downstream motifs have the potential to act at the RNA as well as the DNA level.

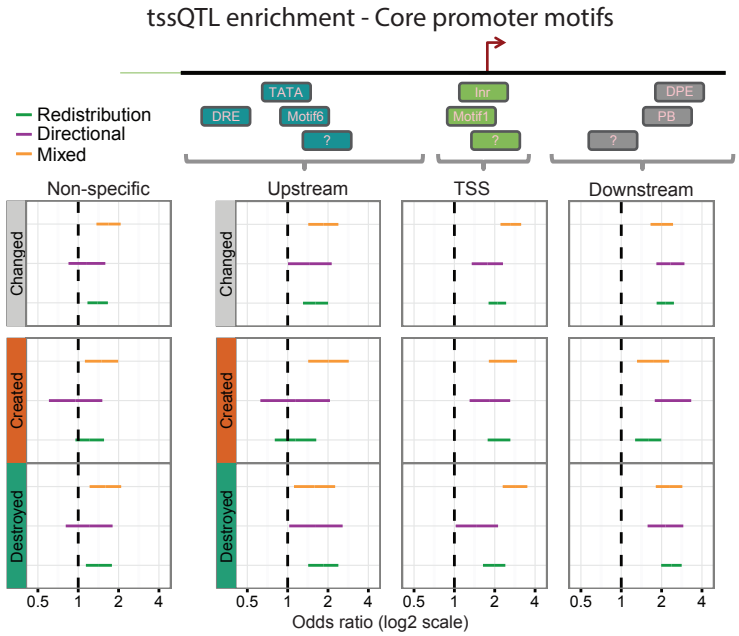


Figure 4.11: Promoter-associated motifs show enrichment for tssQTL.

We included all the motifs that we recovered *de novo*, which include both known and novel core promoter-associated motifs. Creation or destruction is defined according to the major->minor direction. Non-specific motifs include those that are not directional, positioned or discriminative between promoter types. Positioned motifs were separated according to their most frequent position relative to the main promoter TSS. Enrichments (x-axis) are displayed on a log2 scale, and are derived from a logistic regression model (see materials and methods) predicting a variants likelihood of being a QTL based on the relevant gene's expression, the SNP's minor allele frequency, and the relevant feature.

As with our 3' Tag-seq QTL, we decided to examine trends in expression level change, and in shape index change (i.e. the difference in shape index between the two genotypes), for motifs affecting specific motif classes. We first examined tssQTL affecting INR-like motifs, filtering strictly for those that overlapped each QTL's most affected TSS (Fig 4.12). These results show that the creation of an INR like motif tends to cause an increase in shape index. This is exactly as expected, given that the INR motif appears to function as a focal point for transcription initiation. A similar trend is observed for variants disrupting motifs that resemble DPE (or the similar 'pause button' motif – see Hendrix *et al* 2008). Furthermore, broader promoters tend to show smaller changes in expression magnitude than narrower ones (Fig 4.12

c,d). While the small number of cases here does not allow trends to be definitely identified, these results are consistent with a model in which similar mutations acting in broad promoters vs. narrow promoters have different effects, with broad promoters more frequently seeing changes in shape only, and narrow promoters seeing changes in expression as well. These results provide evidence that our tssQTL that our tssQTL are enriched for variants affecting core transcriptional motifs, and point to the differing biology of broad and narrow promoters.

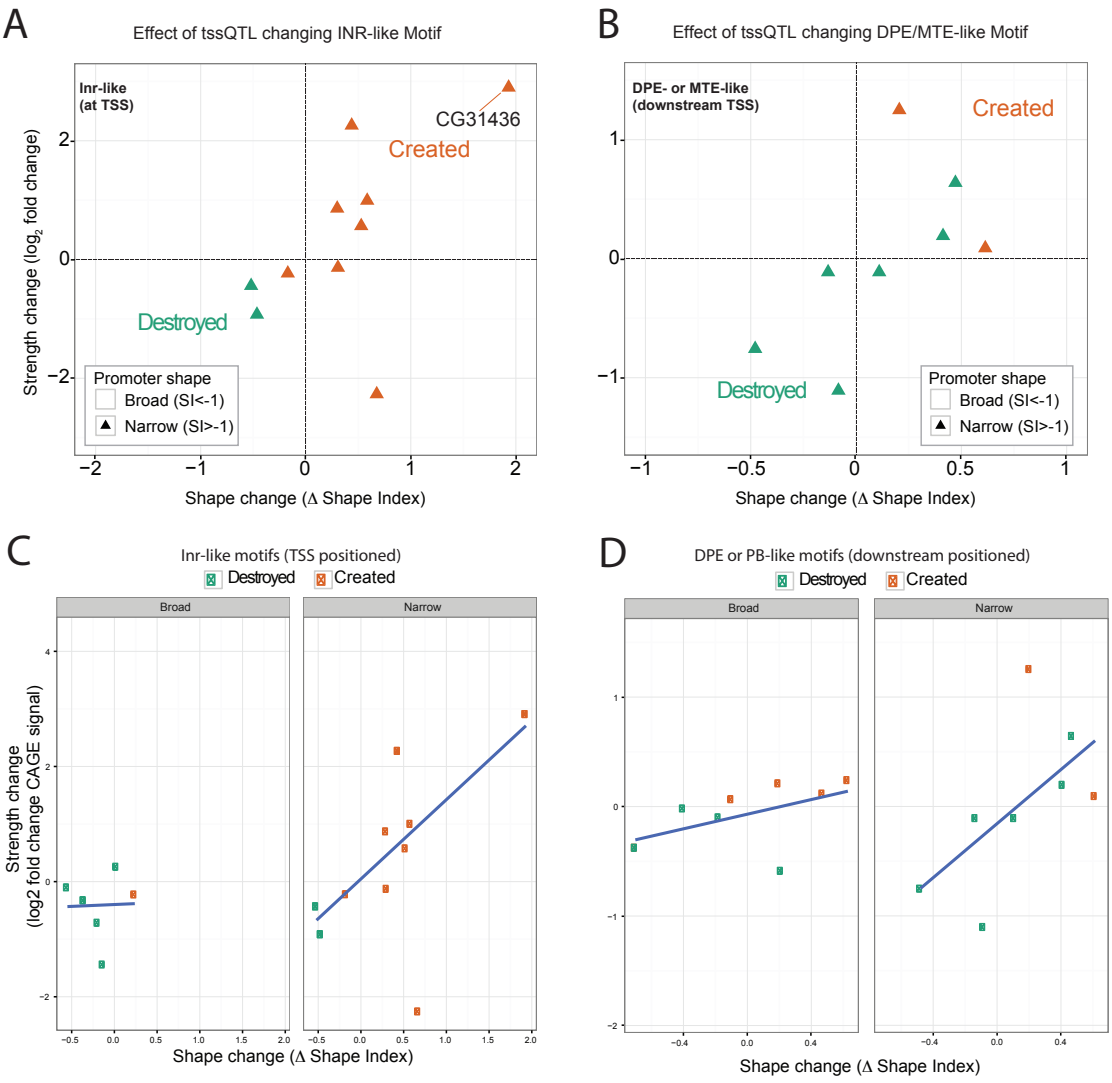


Figure 4.12: Effect of positioned motif turnover on promoter shape, strength

Change in strength (log2 fold-change in mean CAGE signal) and in shape (difference in shape index) between the minor and the major genotypes for a group of QTLs affecting INR-like motifs (A) located at the position of the most affected TSS, or

affecting DPE- and PB-like motifs (B) located 10-40bp downstream of the most affected TSS. Shape of the marks indicates the promoter shape. Promoters are discriminated according to their shape index. The position of the CG31436 gene promoter, the example shown later, is indicated in the "Inr-like" panel. Note that usually an increase in strength is accompanied by an increase in shape index (i.e. narrowing of the promoter) and vice versa. This is especially noticeable for TSS-positioned motifs in the narrowest promoters (A, squares). C,D) Relationship between log2fold expression change (y-axis) and shape index change (x-axis) for INR (C) and DPE/MTE (D) motifs, (as above) but segregated by broad vs. narrow peaks. Note that the correlation in both cases is stronger for narrow peaks – both DPE and INR are motifs associated with narrow peaks.

4.5 Experimental validation of motif changes

To analyze the effects of the TF motif disruptions identified in my analyses of 3' Tag-seq eQTL, we used in vitro luciferase assays (see material and methods), in S2 cells. We chose three motif destruction events for analysis.

The first was a disruption to the motif for Slp1 – a transcriptional repressor that plays a crucial role in early pattern specification in the developing embryo. A variant found at 21% frequency destroys a promoter proximal slp1 motif upstream of the gene CG10306, with a C->T transition disrupting a critical base pair (Fig 4.18a). Consistent with Slp1's predominant role as a repressor, the gene is unregulated in lines with the minor allele.

The second and third disrupted motifs we tested are motifs for the GATA factor Pannier, with both leading to reduced expression of the associated genes, in line with Pannier's role as an activator. One is a variant destroying a promoter proximal Pannier motif upstream of the gene CG17343 (Fig. 4.13) and a second Pannier proximal to the promoter of CG9870 (Fig. 4.14). The second of these QTL is a stage specific QTL which acts only during mid embryogenesis. Each of these motif disruptions was a 'strong change' (score change of 3 or more) and was occupied by the relevant factor. For each of these loci, the promoter proximal region was cloned into S2 cells upstream of a constitutive promoter. For each region, three haplotypes of the locus were transfected – one taken from lines with the major allele sequence for the variant (Major), one from a line with the minor allele (Minor), and one with

the minor allele for the variant introduced into a major genetic background (Maj^{Min}) (Fig 4.13b,Fig 4.14b,d). In all three cases, the (Maj^{Min} genotype showed a significant effect on the constructs expression, and in the direction expected from the QTL's effect, indicating that the intersection of motif analysis, ChIP data, and eQTL analysis can help to more accurately identify causal variants in *Drosophila* . Notably, in the case of our Slp1 construct, transcription from the construct was dependent on the overexpression of Slp1, confirming that motif analysis had accurately identified a causal mechanism behind the eQTL (Fig 4.13 b). Additionally, in all three cases, the minor haplotype also showed a weaker effect than the Maj^{Min}. This indicates that epistatic interactions between motif disruptions and surrounding variants are acting to moderate the effects of the functional variant – selection acts on haplotypes rather than individual variants, so that the haplotypes actually segregating in the population are those with buffering epistatic effects. As such, our results constitute one of the few experimental demonstrations of epistasis in a natural gene regulatory system, and support speculations that interactions between loci, while difficult to identify statistically owing to power issues, may nonetheless play an important role the in the genetics of expression variation.

Epistasis is a phenomenon of great importance not only to our understanding of basic biology, but also to our understanding of current studies in human genetics. Genome wide association studies have consistently found little evidence for epistasis. However there is a crucial distinction to be made between statistical epistasis and biological epistasis – even where the actual mechanisms of variation are epistatic, as experimental evidence suggests they are, we expect low allele frequencies to let an additive model capture most of their contributed variance. Crucially however, we then also expect the apparent additive variance contributed by an allele to change drastically between genetic backgrounds. Our work therefore joins an important body of work collected from studies in model organisms (e.g. Huang *et al* 2012) suggesting that epistatic interactions may be a ubiquitous feature of genetic variation, and one plausible reason for the low cross-study replication rate in human studies. (Mackay and Moore 2014).

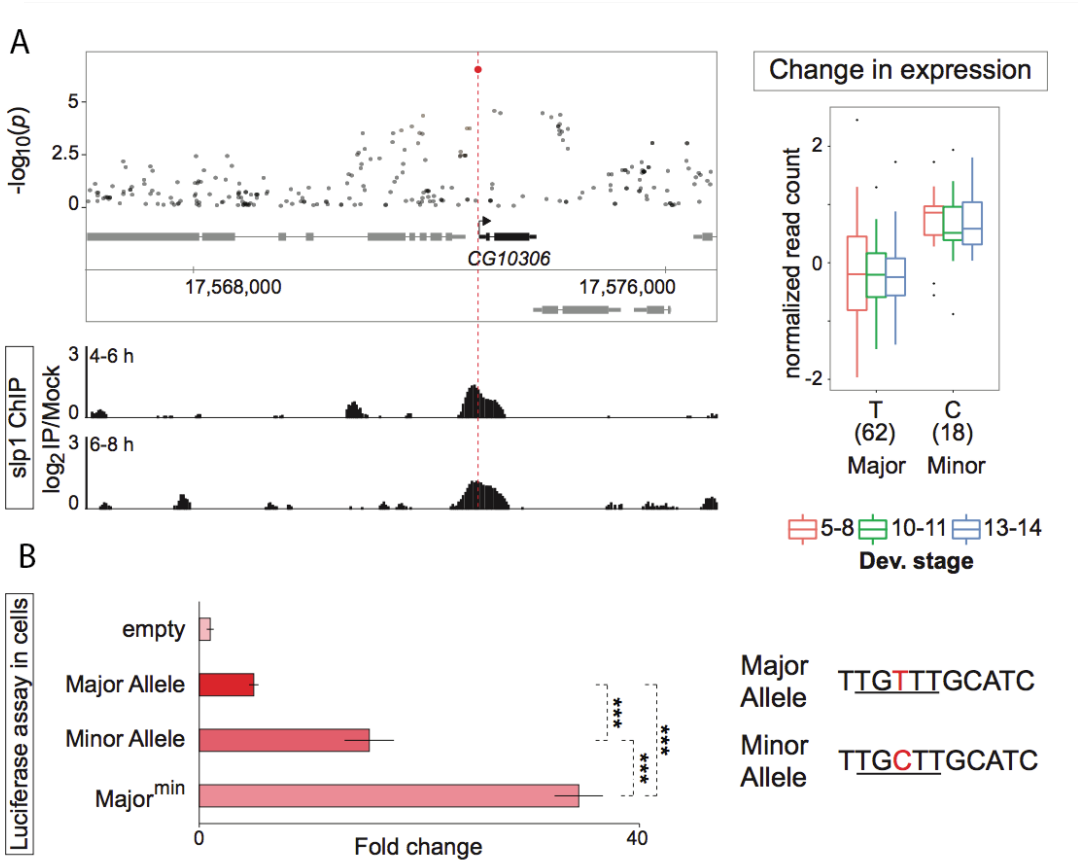


Figure 4.13: 3' QTL at CG10306 shows epistatic buffering of expression levels.

A) Manhattan plot showing unadjusted –log10(P-value) for all tested variants within the extended *CG10306* locus. Red dot indicates the most strongly associated genetic variant. Lower panels show Sloppy paired 1 (Slp1) ChIP at 4-6hr (stages 8/9) and 6-8hr (stages 10-11, matching the middle time-point of the QTL study) of development. The gene's expression is increased in embryos with the minor genotype at all three stages (box-plots, right). B) Luciferase assay of *CG10306* promoter-proximal element, showing 5-fold activity for the major genotype, and an 15.5-fold increase in the minor genotype. When the minor allele variant in Slp1 TFBS is introduced into the major genotype genetic background (Maj ^{Min}), expression is increased to 34.5-fold. ***=pvalue <0.001, t-test. Right, the single bp change in the Slp1 motif in the minor allele.

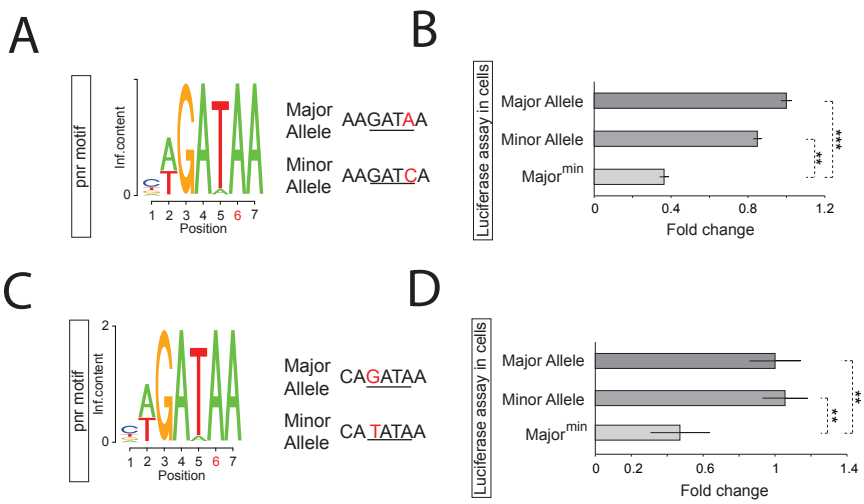


Figure 4.14: Effects of 3' QTL disrupting Pannier motifs show epistatic buffering of expression levels.

A) The single base pair change to the Pannier motif in the promoter proximal element of the *CG17343* locus. B) Luciferase assay investigating mutations at the promoter-proximal element, for the *CG17343* locus. Assay shows 4.6-fold activity for the major genotype, and a 2.9-fold decrease in the minor genotype. When the minor allele variant in TFBS is introduced into the major genotype genetic background (Maj^{Min}), expression is decreased to 1.3-fold. *= pvalue <0.05 ***=pvalue <0.001, t-test. C), as in (A) for the *Eogt* locus. D) Luciferase assay for the *Eogt* 3' element, showing 1.7-fold activity for the major genotype, and 1.8-fold in the minor genotype. Although the GATA site mutation has little effect within the minor genotype, it causes a significant reduction in expression when introduced into the major genotypes background (Maj^{Min}), reducing expression to background levels. **=pvalue <0.01, t-test.

To analyze the effects of variations in promoter-associated motifs causing Direction tssQTL, we elected to use a transgenic reporter assays, similar to those used to analyze the transcription factor motif disruptions in 3' Tag-seq eQTL, but using flow cytometry to obtain measurements for individual cells. Test promoters were cloned upstream of an sfGFP reporter gene, with an mCherry reporter gene coupled to a constitutive reporter on the plasmid as an internal control. For validation, we selected three motif disruptions. The first was a disruption to the INR motif of the gene CG31436, (Fig 4.15). The second was a disruption to the Motif 1 (which fulfills a similar function to INR but is seen in broad peaks) of the gene CG12576 (Fig 4.15). The third is a disruption to a DPE motif found in the promoter of the gene Hn. As with the 3' Tag-seq QTL, we constructed Major, Minor, and (Maj^{Min}) constructs for

each SNP. The results (Fig 4.15) demonstrate that as with our 3' Tag-seq QTL, motif analysis can provide information about which variants are likely to be eQTL. We did not observe the epistatic buffering effects in these mutations that were observed in all the 3' Tag-seq QTL - the minor allele and Maj ^{Min} genotypes are of similar expression values in the minor and Maj^{Min} genotypes. This is plausibly because these mutations affect core transcription initiation motifs, which show a specific grammar that TF motifs generally lack, and might therefore be less easily compensated for by other mutations in other locations. There could also be more tolerance for QTL affecting CAGE levels, if they are less reflective of the final transcript concentration, which could result from QTL affecting abortive transcripts associated with stalled polymerase, more than complete transcript levels. Alternatively, given the small sample size involved, there may not actually be any difference in the levels of epistasis seen in tssQTL vs. 3' Tag-seq QTL.

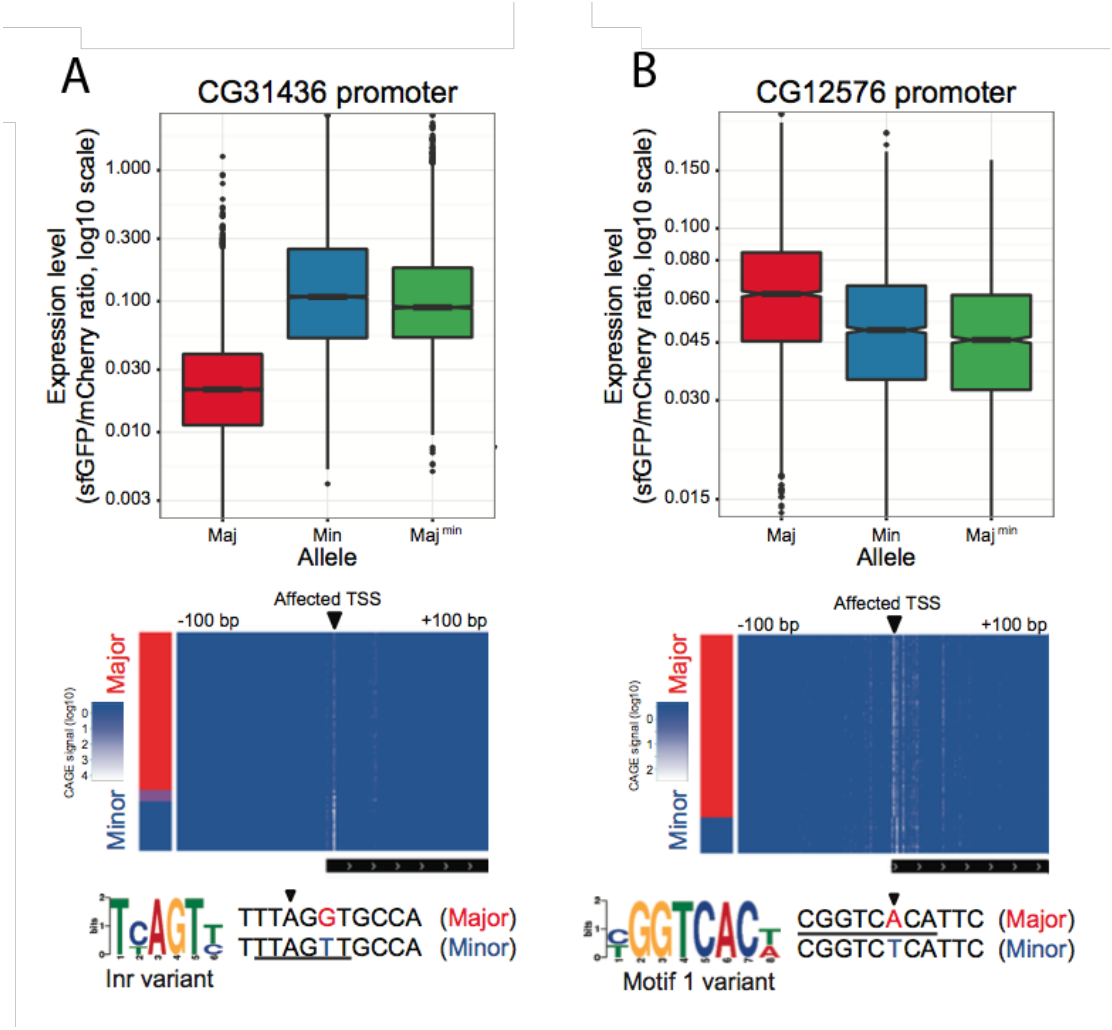


Figure 4.15: tssQTL disrupt promoter-associated motifs.

Examples of motif turnover at the TSS of narrow and broad promoters. A) INR motif creation in a narrow promoter. B) Motif 1 disruption in a broad promoter. Boxplots show expression levels as per experimental validation. Heatmaps show CAGE signal for minor and Major genotype. In both cases, we included natural promoter variants amplified from fly lines representing the major and minor allele (Maj and Min), and also point mutants of the Maj variant with the most significant SNP changed to the minor allele (Maj^{min}). Boxplots show the distribution in the population of single-cell expression values for the different promoter variants in log10 scale. The genotypes differ significantly from one another ($p<0.05$, Wilcoxon rank test) we show below the heatmaps with the raw CAGE reads for each line in a zoomed promoter window, indicating the position of the highest effect (arrowhead). Also, the bases around the likely causative variant of the two variants are shown, together with the affected motif consensus. The created or destroyed motifs in the promoter sequence are underlined, and an arrowhead marks the TSS.

4.6 Relationship between Promoter Shape and tssQTL type

Having observed different trends in the effects of tssQTL in different promoter shape-classes, we moved on to ask about the relationship between shape itself and the presence of tssQTL. Using the shape scores for CAGE windows (used to calculate tssQTL), we find a significant association between shape and the likelihood of having a tssQTL (OR=3.05 p= 2.25 x 10-172, fisher's exact test), primarily due to an increase of Redistribution QTLs. (Fig 4.16a). Reasoning that this could simply be a result of some broad windows including multiple independently regulated TSS, and hence having more opportunity to show tssQTL, we then asked if the same phenomena was visible at the level of CAGE peaks. We first wished to assign QTL status to our CAGE peaks, which could not be done by simple CAGE window-CAGE peak analysis, since many windows overlap multiple clusters. We therefore assigned QTL status to each of the 'main' CAGE peaks from out TSS motif analysis, by taking the location with the strongest effect size for each QTL, and then assigning a QTL to a cluster if it overlapped this location. We observed that broad clusters are also more frequently affected by tssQTL than narrow clusters (odds ratio 2.14-2.52, $p = 3.08 \times 10^{-91}$ Fisher's exact test). Furthermore, we observed that this effect is primarily the result of an increased number of Redistribution QTL affecting broad TSS, (odds ratio 2.24-3.13, $p = 3.94 \times 10^{-31}$) rather than Directional QTL (odds ratio 0.32-0.43, $p = 7.10 \times 10^{-41}$), with mixed tssQTL intermediate between the two (odds ratio 0.91-1.27, p = 0.52).

One simple explanation for the increased number of tssQTL in broad promoters is power. Since broad TSS tend to be more highly expressed, this could cause them to show more detectable tssQTL (Fig 4.2). We also wanted to see the relation between shape and QTL probability independently of the binary "broad" and "narrow" categories, and treat shape as a continuous variable. As shown in Fig 4.16b (open dots), a decrease in tssQTL frequency is observable over the distribution of shape indices. We therefore used a logistic regression approach to model the likelihood of each CAGE Window having a tssQTL, comparing the results of a model incorporating only the total expression of a window, to one incorporating both its total expression and its shape (Fig 4.16b – red vs blue dots). A likelihood ratio test

for the two model shows that the effect of shape on the frequency is independent of expression ($p<2x10-10$), indicating that the increased number of QTL in broad promoters does not result only from increased power to detect their tssQTL. Furthermore, we also find that the relationship between shape and tssQTL frequency is attributable mostly to Redistribution (Fig 4.16 c) QTL, with shape providing little to no extra explanatory power over expression for Directional QTL (Fig 4.16d).

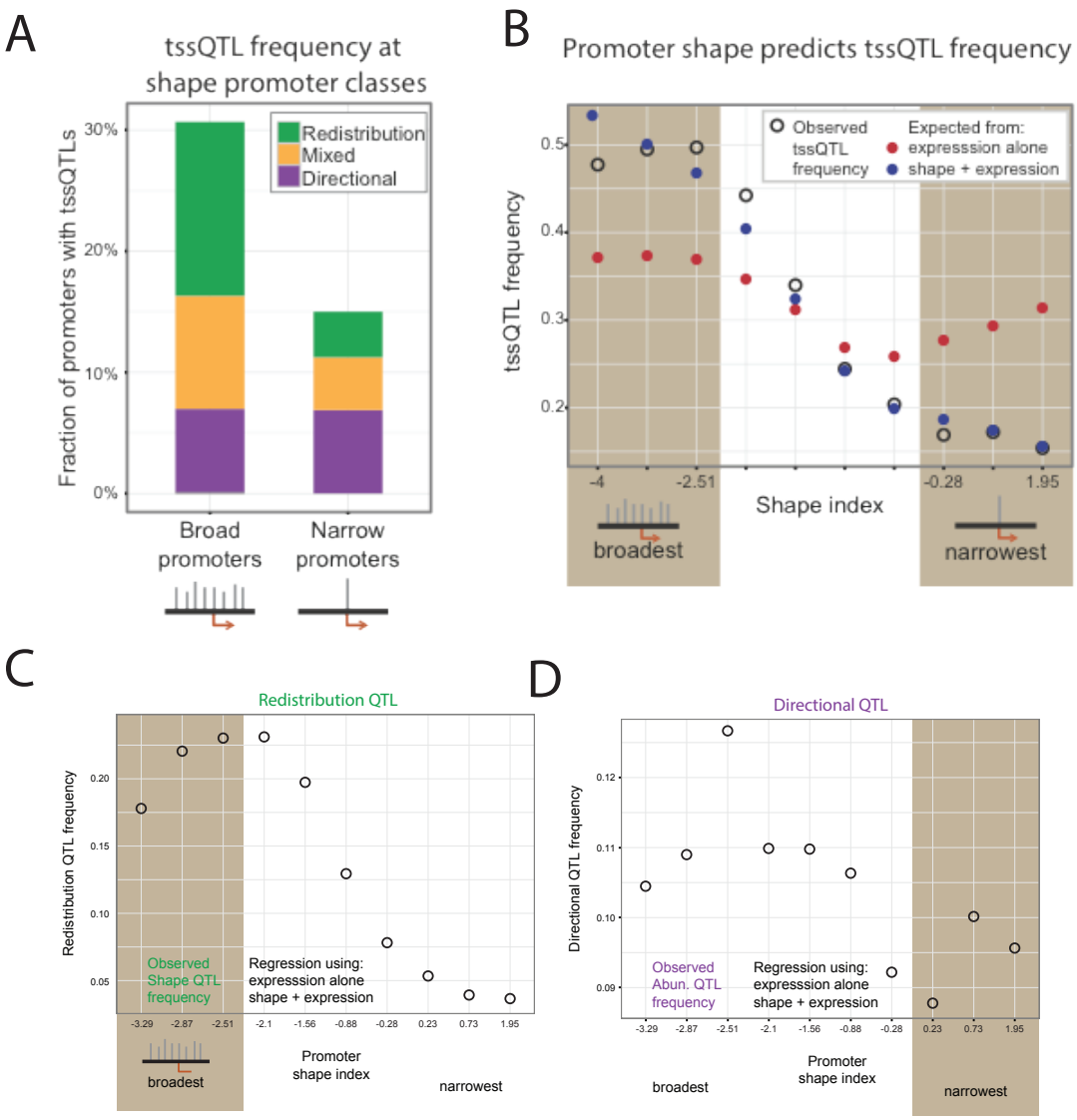


Figure 4.16: Broad promoters show an increased number of Redistribution QTL.

Fraction of individual TSS clusters, either broad or narrow, affected by tssQTL of each class. Broad promoters are more likely to have a tssQTL, and are more frequently affected by shape QTL (green). B) Predicted (filled circles) and observed (open circles) tssQTL frequencies across different shape bins, calculated on the 1kb CAGE windows. Predictions were made with logistic regression models considering only expression level (red dots) or both expression and shape index (blue dots). Shaded regions show the extreme 30% for each shape type. C, D) Same as B but using only Redistribution, or only Directional QTL. Note that the predictive power of shape is low forDirectional QTL

To provide a confirmation of our results that did not rely on CAGE data to classify promoters, we also classified promoters using the Ohler motifs, separating them into five classes (Ohler *et al* 2006): promoters possessing only an INR motif, a TATA box plus and INR motif, an INR motif plus a DPE motif, promoters containing a DRE motif, and promoters containing both an instance of Ohler Motifs 1 and 6. The latter two classes are associated with broad TSS, while the first three are associated with narrow TSS (Rach *et al* 2009). Selecting TSS that fit cleanly into these five classes, i.e. with only those motifs in their expected positions (Fig 4.14a), we analyzed the classes for tssQTL frequency. Consistent with our previous results, we find that motif classes associated with narrow TSS have lower fractions of tssQTL (15.2%, 14.1% and 17.8%) than the classes associated with broad TSS (33.4% and 45.5% for the broad classes), with a 2.1-fold difference between the two ($p = 3.15 \times 10^{-20}$, fisher's exact test). Again, consistent with our previous results, we find that this difference is attributable to Redistribution QTL, with the narrow classes in fact have a higher proportion of Directional QTL (Fig 4.14). We conclude that broad promoters in fact possess more variants affecting transcription initiation patterns.

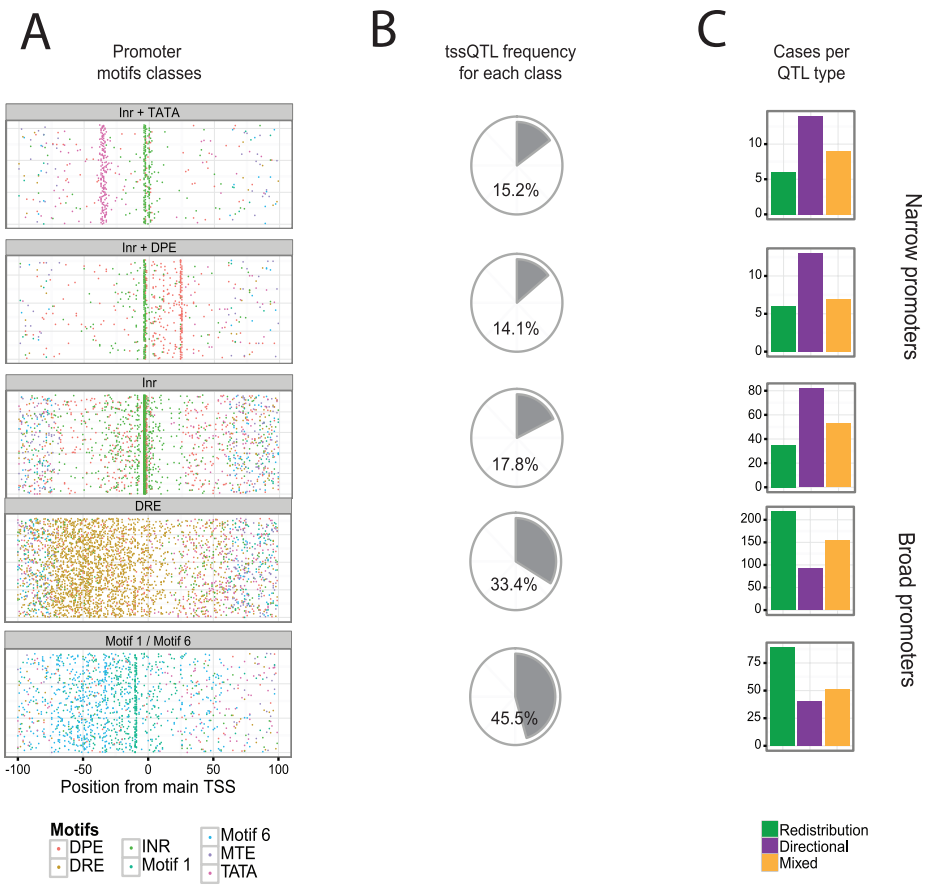


Figure 4.17: Promoter-associated motif classes associated with broad promoters also show increases in Redistribution QTLs

For each promoter class defined by motif content (Ohler, 2006) we show: (A) positioning of motifs, (B) fraction of promoters involved in a tssQTL, and (C) number of cases from each tssQTL type. Only promoters falling unambiguously into one of the six classes (INR+TATA, INR+DPE, INR, DRE, Motif 1+6) were included. Only motifs within functional zones (e.g. +/- 10bp relative to the TSS for INR) for each motif were taken from Ohler 2006. Broad promoter classes are significantly more likely to have tssQTL ($p = 3.15 \times 10^{-20}$, Fisher's exact test).

An obvious explanation for this phenomenon is that broad promoters present a larger mutational target size, because of their greater width. To determine if this was the case, we grouped our CAGE windows into ten quantiles by shape, and used INSIGHT to obtain the proportion of sites under selection within each shape-bin. We find that this proportion is between 0.61 and 0.65 for all promoter shape bins (Fig 4.18a). Note that all windows first had coding sequence removed, since this would

otherwise present a major confounder. Incorporating the fraction of functional sites into the logistic regression used to predict a windows tssQTL frequency gives only a minor improvement to the models predictivety, indicating that an increased fraction of functional base pairs does not explain the increased frequency of tssQTL seen in broader TSS (Fig 4.18b).

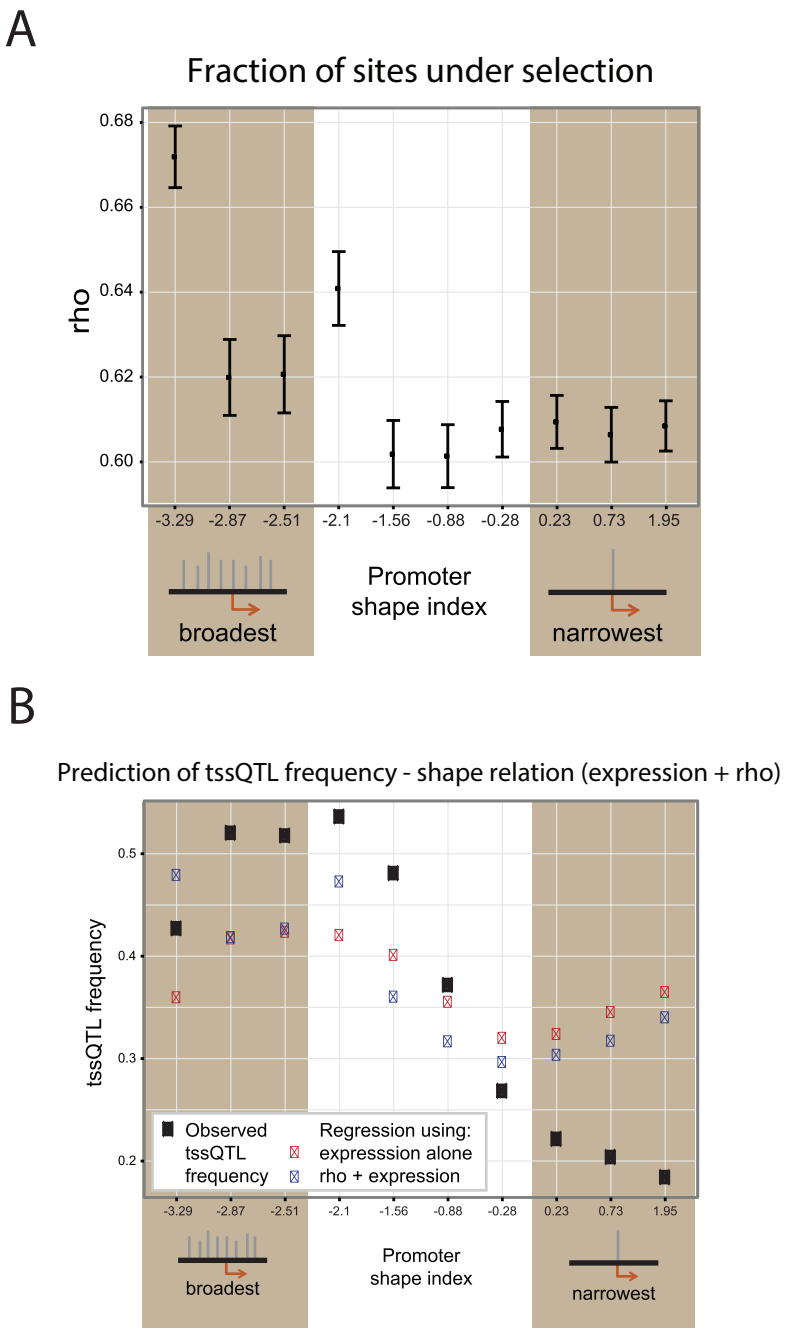


Figure 4.18: Differences in number of sites under selection does not explain relationship between tssQTL frequency and shape.

A) Estimates of rho (fraction of sites under constraint, y-axis) for the 1kb CAGE windows used to call eQTL, binned according to their shape index (x-axis). The broadest peaks have a higher fraction of functional sites, probably due to their increased likelihood of being found internal to genes. B) Predicted (filled circles) and observed (open circles) tssQTL plotted over same shape bins. The rho value for its shape bin was assigned to each shape bin and used as a covariate in a logistic

regression over all CAGE windows. Rho is only slightly predictive of the tssQTL likelihood, compared to shape.

In effect, broad promoters appear to be subject to an additional type of variation. This variation affects the shape of TSS rather than their overall output. One explanation for this is that the presence of multiple functional initiation sites at a TSS buffers the effects of mutations at individual sites, such that they are less subject to negative selection. In support of this, while minor allele frequencies for both classes of tssQTL are similar, the frequency of the derived allele for Redistribution QTL is significantly higher than for Redistribution QTL ($p = 0.04655$, one tail Welch t-test), consistent with a scenario in which these mutations, once they arise, are less constrained, and thus more likely to reach high frequencies.

In examining the evolutionary properties of broad promoters using INSIGHT, we also observed another surprising difference - broad promoters show significantly (~ 3fold comparing narrowest to broadest) higher levels of adaptive substitution (Fig 4.19). To confirm that this result was not an artifact caused by the assumptions of INSIGHT's model (which could result from, for instance, difference in mutation rates between broad promoters and the control regions used by INSIGHT caused by broad promoters being more exposed to mutation). We also plotted some basic statistics for the genetic variation in the 10 shape bins. We observe that the absolute level of divergence, and fraction of non-ancestral polymorphisms, is also higher in broad promoters (Fig 4.20) and that level of polymorphism, and ratio of rare to common polymorphism, is lower. This is inconsistent with a model in which increased divergence at broad promoters is simply due to a higher levels of polymorphism, supporting the conclusion that that there is an increased level of adaptive substitution within these promoters.

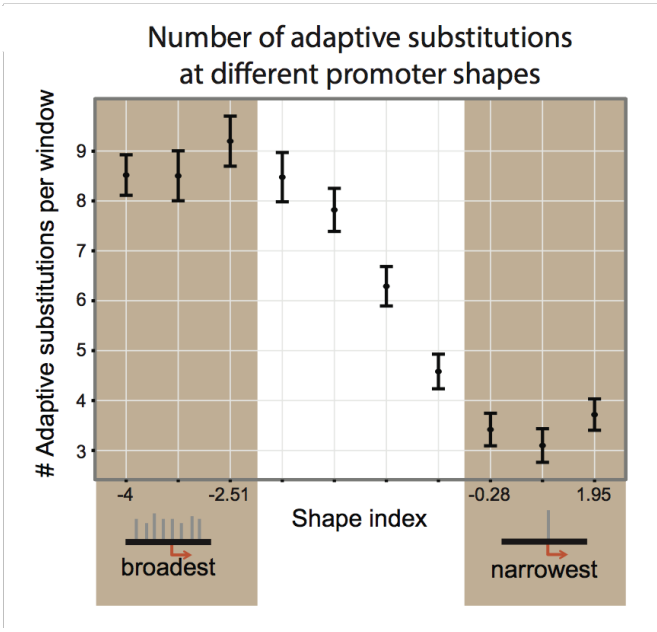


Figure 4.19: Broad promoters show an increased number of adaptive substitutions

Estimated number of adaptive substitutions as per INSIGHT analysis (Arbiza *et al* 2013) for ten groups of CAGE windows, binned by shape index (x-axis). Plot indicates the estimated number of adaptive substitutions per bin (y-axis). Narrower CAGE windows have fewer adaptive substitutions.

We wondered if the phenomena of increased adaptive substitution at broad promoters could be linked to increased variation permitted by their distributed architecture. Such a link would be puzzling, if there were selective disadvantage to mutations affecting promoter shape. We therefore decided to experimentally investigate our Redistribution QTL to determine if they had effects on gene expression that were not apparent at the population level.

4.7 Redistribution QTL affect Expression noise, and are buffered by epistasis

We also wished to experimentally investigate the effects of Redistribution QTL. Since Redistribution QTL do not affect the overall level of expression, we reasoned that they might affect other properties of the promoter, such as the cell-to-cell variation in transcriptional output, or 'expression noise'. To assay this, we again used the promoter reporter system and examined the variance between thousands of

individual S2 cells. We selected five cases with Redistribution or mixed effects where a clear change in shape is seen, and where a single SNP showed a P-value 10 fold or greater than all others. We selected three Redistribution QTL with motif disruptions: the variant affecting CG17802 shows disruption to an motif which resembles the E-box motif, while the variant affecting Spt6 affects a positioned motif enriched in broad promoters, AAHAAW, which has both positional enrichment and a modest resemblance to Ohler motif 6. The variant affecting CG11210 affects RTGYA, a short, low information content motif discovered in broad promoters that lacks positional enrichment or resemblance to any other motif. We also included variants affecting CG7927 and CG2469, which, although they do not affect detectable motifs, nevertheless have strong affects on promoter shape. The variant affecting CG2469, however, interrupts a poly dT tract preceded by a long dA tract. Such sequences are known to disfavor nucleosome positioning, and this could be a causal mechanism for the shape change in this tssQTL - which does not affect overall expression level. For each promoter we again constructed Minor, Major and Maj^{Min} genotypes (Fig 4.20), and assayed both the magnitude and variance of expression.

In each case, there is some change in overall level of expression. We note that there is no detectable 'buffering' epistasis for expression level, except in the case of CG17802, which, significantly, also shows the largest change in expression level of the five QTL. In the other four, expression levels remain static or diverge slightly further from the major allele when comparing Maj^{Min} to Minor, suggesting that smaller changes in expression levels might be tolerated and therefore appear in haplotypes without buffering mutations. In four of five cases however, we detected a significant increase in expression noise in the Maj^{Min} genotype, as compared to either Major or Minor. In the exception - CG11210 – the major genotype is noisier than either Minor or Maj^{Min}, the former is still less noisy than the latter. This promoter has a disruption to a short motif without positional enrichment, which may mean that its lack of effect on expression noise is simply a result of it not affecting core transcriptional machinery in the same way as the others. We also note that the change in variance is not simply a function of expression level, since genotypes increasing or decreasing overall levels exist that increase noise.

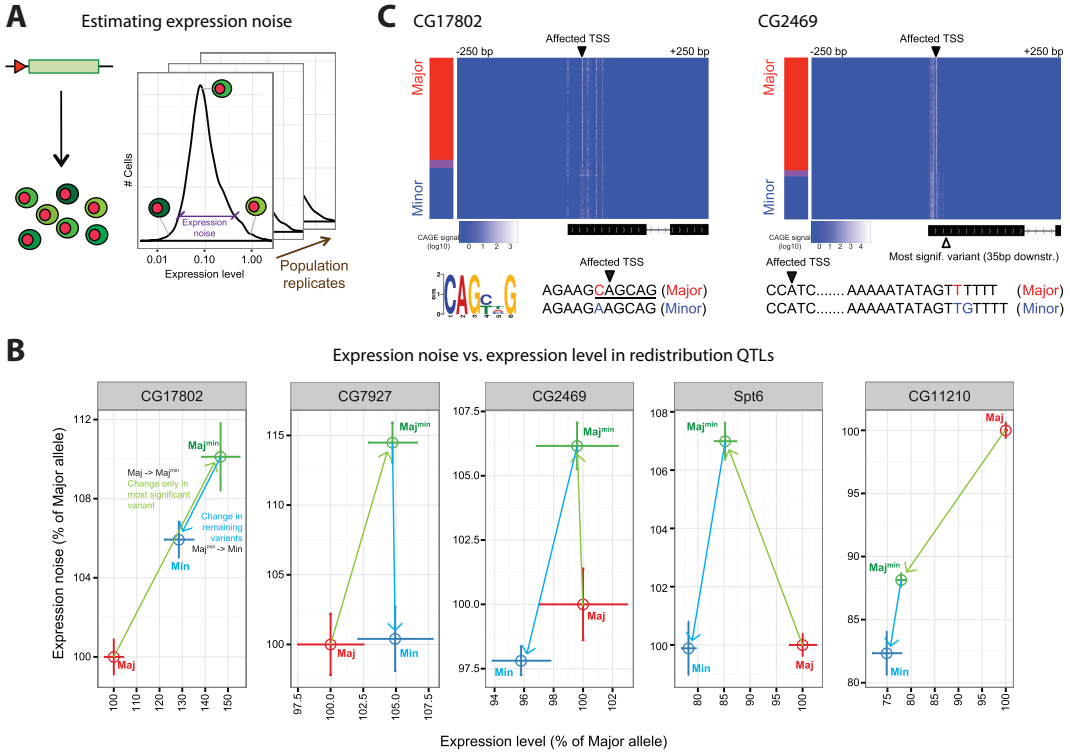


Figure 4.20: Expression noise resulting from shape transitions caused by disruption to promoter-associated motifs is buffered by epistatic effects.

A) Schematic view of how differences in expression noise are measured. We measured noise as cell-to-cell variation (median absolute deviation from the population median) on the expression values of a reporter driven by the variants of the indicated promoters (same reporter as Fig. 4.15). B) Expression noise resulting from shape transitions is ameliorated by epistatic effects. Expression and noise for the transitions Maj -> Maj^{min} (blue arrow) and Maj^{min} -> Min (green arrow). Maj^{min}.= lead SNP engineering into the sampled major haplotype. While in 4/5 cases, the mutant variant (Maj^{min}) showed increased noise, in all cases the presence of the other variants in the Min haplotype reduced the noise from the Maj^{min} mutant haplotype. C) *Examples of shape QTL with increased noise:* Heatmaps showing CAGE signal in a zoomed promoter window for two examples. For *CG17802*, a novel TSS-positioned motif (underlined) is interrupted by the SNP. For *CG2469*, a G insertion interrupts a dT-tract, although no specific motif is predicted to change. Both dispersions and mean expression values are defined from three biological replicates (errorbars represent +/- 1 SE) with > 15,000 cells measured per replicate.

4.8 Analysis of expression constructs reveals candidates for epistatic interactions.

The results of our experimental validations of both 3' Tag-seq eQTL and tssQTL suggested that eQTL causal mutations might frequently occur in 'permissive haplotypes' – i.e. in linkage to variants that decrease the sum effect of the haplotype on gene expression. Given the differences between Maj^{min} and the minor allele, we reasoned that other variants should be present in the constructs differing between the major and minor lines used, other than the variant targeted for mutation. These variants would then be candidates for the observed haplotype effects. To investigate this hypothesis, we plotted the construct loci in each of the 3' Tag-seq eQTL constructs (Fig 4.21 A-C) and tssQTL constructs (Fig 4.21 D-J) used in our studies. We found that in each construct, at least 2 other variants were present that could potentially account for the haplotype effects. The number of DGRP lines used in our study did not permit secondary effects to be detected statistically in the original data, even when narrowing the search space down to these candidate variants. We also did not observe any obviously compensatory motif changes (i.e. cases in which the motif destroyed by our original targeted mutation was created by one of the candidate secondary variants within the construct). Thus, while the other variants present in these motifs are evidently effecting gene expression, given the observed difference between the minor and Maj^{min} genotypes, we are unable to determine an obvious mechanism. Given that a large fraction of our eQTL do not disrupt our high quality motifs, this is not necessarily surprising – these other motifs could be compensating for observed motif disruptions by changes to any of many factors for which we could not find high quality motifs.

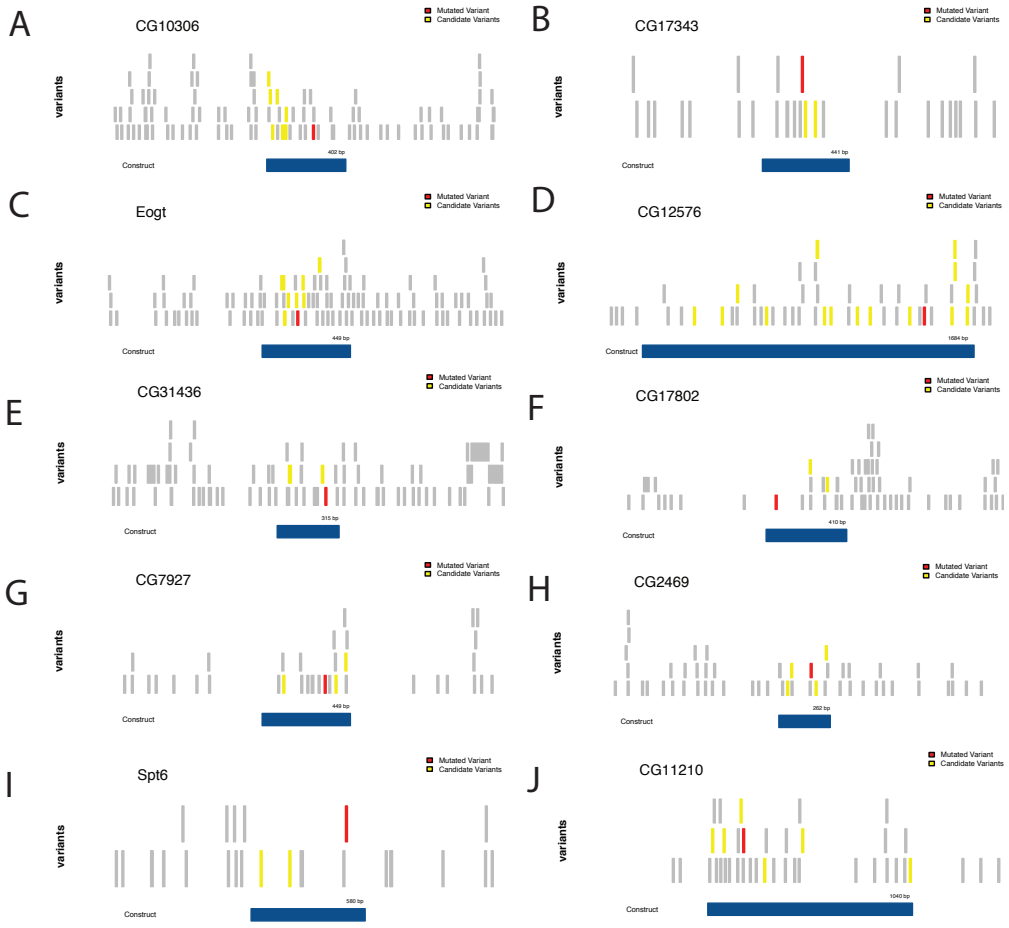


Figure 4.21: Mutated variants and candidate interacting variants in expression construct loci.

Shown are the positions of variants within each expression construct tested. Grey bars mark the positions that are polymorphic within the DGRP collection. Blue bars represent the region used to create the construct. Red bars indicate the variant that was mutated to create the Maj^{Min} genotype for each construct, while the yellow bars indicate other variants that also vary between the minor and major genotype tested, and which are therefore candidate interacting variants. At least two such variants are present in all cases, meaning multiple variants may contribute to the phenotype. For the 3' Tag-seq eQTL constructs (A-C) the candidate variants did not alter the relevant TF motifs altered by the main variant (i.e. slp1 (A) or Pannier (B, C) motifs). For the tssQTL constructs that altered motifs, (D, E, F, I, J) candidate motifs did not alter the relevant promoter-associated motifs.

4.9 Discussion

4.9.1 Intersecting functional genomic and eQTL data

In this chapter, I have discussed my analysis of the location, properties and mechanisms behind variants affecting gene expression, both at the level of transcription initiation and polyadenylation. One dominating trend in the variants associated with variation in gene expression that I have analyzed is their location – my analysis agrees with studies in other organisms indicated that functional variants influencing gene expression levels tend to cluster around TSS (e.g. Kwan *et al* 2008, Pickrell *et al* 2010,Veyrieras et al 2008). 3' Tag-seq QTLs also cluster around TES, which is no doubt reflective of their biology, however mapping bias may also play a role. While great care was taken in both QTL sets to avoid mapping bias (see materials and methods) lingering heterozygosity and/or genotyping errors cannot be ruled out without additional experiments. The fact that that clustering around TSS occurs for both tssQTL and 3' Tag-seq QTL however indicates that mapping biases are unlikely to be the only cause of their spatially clustered distributions.

My findings of enrichment for QTL within functional elements such as DNase hypersensitivity sites, and transcription factor binding sites, are in agreement with findings in human (Gaffney *et al* 2012). The relatively modest predictive power of most motifs, and CRM associated features, such as DHS, are likely results of several factors. One is functional density of the *Drosophila* genome, with 13.1 % of the genome being hypersensitive (Thomas *et al* 2011) and these sites biased towards genes, and hence our tested variants, there is relatively little non functional sequence for annotations like DHS to provide enrichment over. Also a factor is our relatively small sample size – given just 80 lines, alleles with strong effects may simply not be present in high frequencies in our data. Instead, our data will be biased towards that portion of genetic variation that has low fitness costs, and to the extent that features like CRMs and TFBS are essential, QTLs affecting them may be absent from our data. Another factor is the simple fact that so much of the regulatory genome is poorly understood. The relatively small number of variants that I could identify as disrupting TFBS is no doubt partially a result of the small number of High quality PWMs I was able to identify. Only 36 of the estimated 708

(Hammonds *et al* 2013) were included in my dataset, however these are far from a random sample of all factors, and will most likely be biased towards the most ubiquitous, important factors, since these are most likely to be identified by genetic screens and studied. Presumably as the quantity of ChIP and PWM data available grows, the number of variants with plausible candidate mechanisms will climb. Until then, such variants potentially indicate undiscovered regulatory mechanisms, and provide a promising set of candidates for experiments. My results underscore the fact that while a large quantity of functional genomics data now exists in the literature, its quality cannot be taken for granted.

In contrast to the destruction of functional sites, mutations creating functional binding sites have received comparatively little attention. My results indicate however, that many functional variants are plausible gain of function mutations. Future QTL studies should address the possibility that functional variants create, rather than destroy, functional motifs.

The absence of a correlation between PWM score change and effect size is likely a result of several confounding factors. Firstly, there is evidence (Spivakov *et al* 2012) that high frequency variants that strongly disrupt PWM motifs are more rare in *Drosophila* than in human, indicating more effective selection in *Drosophila* . This may account for both the relative paucity of QTL in any given motif, and the lack of correlation with effect size (stronger constraint would mean our observed PWM changes are more highly enriched for those TFBS which are non functional false positives). Second, recent evidence has shown that PWM score alone is a relatively poor guide to the effect of a functional variant - e.g. Das *et al* (2016), were able to identify correlations between PWM score change and population level phenotypic variation only by integrating PWM score into sophisticated mixture model incorporating DNase footprinting, which can precisely identify which motif matches are functional. In the absence of high quality DNase footprinting data, we are unable to implement Das et al's model. Third, much transcription factor binding seems non-functional (Cusanovich *et al* 2014) and even where functional binding is present, it is not wholly dependent on the factor's motif (Junion *et al* 2013). Finally, our QTL are distributed amongst a large number of different PWM, which likely have differing biochemistry – many factors can function as both repressors and activators for

instance (e.g. Rembold *et al* 2014). A study focusing on the binding of a single factor such as CTCF (Ding *et al* 2014) could therefore be expected to identify trends we could not.

Many of the studies into genetic variants affecting gene expression have focused on the role of mutations that interrupt transcription factor occupancy (Degner *et al* 2012, Moyerbrailean *et al* 2016). Gene expression, however, is regulated at many levels, with recent studies pointing to an underappreciated role for splicing as as a key mechanism in human disease genetics (Xiong *et al* 2015) and as an important player in the evolution of gene expression (Brawand *et al* 2011). It is therefore striking that we were able to identify so much variation in 3' Isoform usage, and 3' UTR length. The frequent disruption of CPSF motifs in these functional variants strongly suggests that although these motifs are crucial to the regulation of polyadenylation, the process can often still proceed, albeit with some effect on mRNA levels, when they are disrupted. Other motifs disrupted by 3i QTL, such as those for Elav, and Tra2, which have not previously been characterized as involved in polyadenylation, are also interesting. It is possible that these motifs represent motifs for other factors with similar binding preferences, or that the factors themselves have uncharacterized roles in polyadenylation.

Also striking is the enrichment of various promoter-associated motifs for tssQTL since intraspecific variation within promoter sequences has not been well explored, which is in part because of how conserved such sequences are. We were able to experimentally validate the functionality of some promoter-associated motif disruptions (4.15,4.20). Our findings that many other promoter-associated motifs, besides the well known Ohler motifs, carry preferences for broad or narrow promoters, and are positionally enriched, is particularly interesting. We were able to demonstrate that the INR motif for instance, has context dependent effects on promoter shape. The question of whether our other motifs also have context dependent effects, could be answered by experimentally using reporter constructs, and may yield novel insights into the mechanisms determining promoter shape. Such studies will likely need to treat motifs as components of a promoter architecture (e.g.Mitre *et al* 2016) rather than isolated elements – as is illustrated by the context dependent effect of mutations in INR and DPE (Fig 4.12).

My analysis was performed separately from eQTL calling itself, and drew on methods used by others to integrate functional genomic data with eQTL data post-hoc (Brown *et al* 2013). More recently, efforts have been made to construct models that integrate genomic data into the eQTL calling process itself in a joint Bayesian framework (Das *et al* 2016), in which the functional enrichment of elements such as DHS are estimated as parameters of the model. Such models will provide more accurate estimates of the relative importance of different features, however they will also tend to down weight and exclude functional variants that are not within these elements – and may therefore exclude instances of novel biology.

4.9.2 Promoter shape and genetic variation

Our discovery of increased Redistribution tssQTL in broad promoters is novel and further adds to the growing literature on the distinct nature of the regulatory programs affecting broad and narrow genes (e.g. Hoskins *et al* 2011, Rach *et al* 2009, Carninci *et al* 2006). Why would broad promoters tend to show more Redistribution tssQTL? A simple answer is that their many different TSS sites can compensate for one another, such that a mutation affecting one simply allows others to increase in strength. This robustness may then explain the increased positive selection seen in broad promoters. With positive selection an apparently ubiquitous feature of the Drosphila genome, broad, robust promoters could be less constrained, and hence more likely to diverge from ancestral sequences under positive selection.

Evidence suggests that in general, the mutational variance (i.e. the variation in the phenotype resulting from mutation prior to any natural selection) of expression noise is higher than that of expression itself (Metzger *et al* 2014). This could provide an explanation for why Redistribution QTL so often affect expression noise. Our results suggest that 'noisy' alleles in metazoans are typically found within haplotypes with reduced noise levels compared to possible recombinant haplotypes. Because of linkage disequilibrium, positive selection acts on combinations of variants rather than individual ones, and so combinations of variants whose combined effects on fitness are weaker will be more likely to rise in frequency, even where the

variant's individual effects might be severe. One obvious question is the order in which mutations appear - one possibility is that mutations with fitness effects rise in frequency due to subsequent mutations that compensate for their effects. However a more likely explanation, which does not involve the persistence a lower fitness intermediate, is that 'buffering' mutations arise first, such that they create a permissive haplotype in which further mutations can arise which would have been subject to more purifying selection in the ancestral genotype.

Many further questions are suggested by these results – is the mutational variance of expression noise in fact the same in broad and narrow metazoan promoters? Why do broad promoters not show a larger proportion of functional sites if they bind a larger number of transcription factors? Why, mechanistically, do promoters which bind many factors also tend to show widely distributed TSS? A tempting answer to this latter question is that different factors could control different sets of TSS sites, however to our knowledge no variance in TSS site usage has been observed between tissues that would suggest this (FANTOM Consortium *et al* 2014). As our understanding of the underlying logic of transcriptional promoters improves, it should be possible to answer these questions via, for instance, high throughput cell culture assays of many promoter variants, and detailed genetic dissection of different promoter types, along with their response to transcription factor concentrations.

Another striking finding, the high frequency of epistasis observed in our study, joins an important body of literature (e.g. Huang *et al* 2012), which points to the ubiquity of biological epistasis in regulatory DNA. Much of this work focuses on epistasis between physically distant loci, which is more tractable since distal loci will not be in LD, permitting easier analysis of all possible variant combinations. However, anecdotal evidence also suggests that epistasis within regulatory regions may be common has been observed in several targeted studies of human regulatory DNA (Babbitt *et al* 2010,Myers *et al* 2007). *Drosophila* offers an excellent system for analysis of within-CRM epistasis because haplotype blocks are relatively small, and *in vivo* assays can easily test different haplotypes for tissue specific activity. The epistatic effects that we have identified are therefore promising candidates for analysis. Further study should be able to characterize their mechanisms of action,

and yield insights into the mechanisms of gene regulation, and the nature of its variation between individuals, and species. These insights will be particularly interesting with regard to human association studies.

The mostly additive nature of genetic variance in natural populations (Hill *et al* 2008) presents, at first glance, a paradox. Complex organisms are composed of complex, interconnected, non-linear networks, from the biochemical networks controlling cell signaling, regulation and metabolism to the networks of cell-cell interaction governing development and the nervous system. A wealth of experimental evidence supports this, and decades of genetics have demonstrated how frequently the effects of mutations are dependent on other mutations (e.g. Clark and Wang 1997, Elena and Lenski 1997). Furthermore this dependence is not limited to mutations with dramatic phenotypes – QTL (Bloom *et al* 2015) and introgression (Spiezio *et al* 2012) studies demonstrate epistasis even for subtle phenotypes.

Why then can additive models of inheritance so accurately capture the patterns of variation in natural populations? The answer is that allele frequencies for genes with any effect are typically quite low. This is true even of variants usually described as 'common' in the literature. To the extent that this is true, even variants whose effects are biologically dependent on those of others will contribute mainly additive variance (Mackay *et al* 2014). Thus, a distinction must be made between *biological epistasis* in individuals, and *statistical epistasis* in populations. The former does not always imply the latter. Change the allele frequencies of other, interacting alleles however, and the apparent additive effect of an allele can change radically. Recent studies in *Drosophila* (Huang *et al* 2012) are consistent with epistasis being common, and there is no reason to believe that human populations should be different, aside from their population size and linkage disequilibrium.

5 Conclusions

As I write this thesis, the fields of functional genomics and genetics (or more particularly, genome wide association studies) are converging. The time is now coming when the molecular mechanisms that underlie eQTL and GWAS hits can at last be characterized. This convergence will require both improvements to genetics methods, and a more complete understanding of the genome's function.

During my thesis I have furthered our knowledge of the developing *D. melanogaster* transcriptome, by describing the Transcription Start Sites and 3' UTR regions that appear during development, as well as the transcription of *D. melanogaster* enhancers. These three aspects of the transcriptome all represent frontiers of our understanding.

The CAGE data that I have used to delineate TSS in the developing embryo will serve as a useful resource to those studying the nature of transcriptional initiation. Already, in Chapter 4, I have been able to explore the properties of variants affecting promoter shape, and demonstrate context dependent effects due to the disruption of characterized promoter-associated motifs, within natural populations. I have also been able to build on the literature characterizing differences between broad and narrow promoters, showing that they are subject to different types of genetic variation and patterns of natural selection, as well as identify a set of sequence motifs, many of which show different frequency in broad vs. narrow promoters, and which could allow study of the underlying mechanisms separating broad from narrow promoters.

Our analysis of eQTL affecting polyadenylation sites and their motifs also comes at an interesting time, in which splicing, and isoform usage, and 3' UTR biology (Xiang *et al* 2015, Smibert *et al* 2012) are emerging as important parts of metazoan gene regulatory programs. We have identified sequence motifs that could play novel roles in the 3' polyadenylation process, and shown that variation within these motifs is responsible for some natural variation in the usage of polyadenylation sites within D.*melanogaster*, and thus UTR length.

My characterization of enhancer transcription also lays the groundwork for the study of this phenomenon in *Drosophila* . By establishing the presence of eRNA in *Drosophila* , and the similar nature of the phenomenon in human and Fly, my work suggests that the phenomenon is an ancient feature of metazoan genomes – and it is tempting to speculate that it emerges from the basically 'leaky' nature of transcription clearly in evidence even in yeast (Johnson *et al* 2005). With closely related species of *Drosophila* available, it should be possible to characterize the evolution of eRNA at orthologous enhancers. This approach could offer novel insights into whether, and how often, enhancer transcription is functional. Ultimately however, our understanding of the phenomenon will require functional dissection – whether enhancer activity and promoter activity at a locus can be affected independently by mutation, and if not, the nature of the interaction between these two functions, will allow us to understand why they so often co-exist in the same piece of DNA. Such experiments should be easy to carry out in *Drosophila* – knowledge of promoter-associated motifs, if they are found within transcribed enhancers, could provide targets for mutation, and if they are absent, then a 'brute force' approach using high throughput promoter assays could be used, or one based on comparisons between species, which could reveal sequence differences corresponding to differences in enhancer transcription.

Our understanding of enhancers will be critical to the consilience of functional genomics and genetics. It has long been theorized that mutations affecting TFBS and enhancers could be important players in evolution, because the mutations which affect them are likely to be both co-dominant, and less pleiotropic than those affecting protein coding genes (Wray 2007). In my thesis I extend work showing selection in noncoding region (e.g. Andolfatto 2005) and in selected *Drosophila* TFBS to show that that positive selection is present in many *Drosophila* TFBS and ChIP bound regions, strengthening the evidence that in adaptive evolution in gene regulatory regions is a mechanism of evolution.

We have also been able to show epistasis between regulatory mutations in *Drosophila* , and propose that such epistasis is likely ubiquitous in metazoan populations. The ubiquity of epistasis presents a challenge, because it so dramatically broadens the search space for genetic interactions. Even if only pairwise

interactions are considered, testing them all is a significant demand on CPU cycles, let alone on the quantity of data required. In fact, since the human genome is in such heavy linkage disequilibrium, there may not exist sufficient genetic data to test all possible interactions (barring, for instance, the creation of more in cell culture experiments). It will therefore be necessary to inform our inferences about epistasis with additional data, for instance by reducing the dimensionality of the problem by grouping variants with similar effects, or genes with similar function, or by incorporating prior information from cell culture studies and model organisms, or from biochemical models.

One example of such dimensionality reduction that is already possible is the grouping together of coding mutations within genes (Auer *et al* 2015) used in rare variant association studies, which is possible because coding mutations have well understood function. In principle, similar use could be made of information about regulatory mutations, if we could make reasonable predictions about function. Such efforts however will ultimately be limited by our understanding of gene regulation.

Ultimately, if we wish to understand the genotype phenotype map, we will need to understand the regulatory genome. Understanding how one genome can generate such a bewildering complexity of transcriptomes will be key to understanding how we differ from each other, and from other species. The convergence of functional genomics and genetics will revolutionize our understanding of both. The work I have detailed here represents part of our first steps towards this revolution.

6 Methods

The analyses in my PhD were carried out in collaboration both with experimentalists and other bioinformaticians. Here, for clarity and to aid in the interpretation of my own work, I provide details on their methods, as well as my own. Sections describing others work are indicated by a name in parenthesis after the section heading.

6.1 CAGE and 3' Tag-seq data collection and processing

6.1.1 Embryo collections, RNA extraction, CAGE and 3'Tag-seq preparation (Enrico Cavanno)

Freshly eclosed adults were placed in embryo collection vials, with standard apple cap plates. After at least three 1 hour pre-lays, the flies were allowed to lay for 2 hours, after which the embryos were aged to the appropriate time-point. Embryos were then dechorionated using 50% bleach and snap frozen in liquid nitrogen. As we are interested in variation in gene expression under genetic control, rather than stochastic differences between individuals, RNA was isolated from a pool of 100 embryos collected from the same adults. Embryos were homogenized using a Cordless Motor for Pellet Mix and pestels (VWR) on ice. RNA was extracted in TRIzol®LS (Life Technologies), digested with RNase free DNase I (Roche) and purified with the RNeasy mini kit (QIAGEN) according to the manufacturers' recommendations.

The DGRP inbred *Drosophila* lines were derived from a single, natural population in Raleigh, NC (USA) followed by more than twenty generations of controlled inbreeding (Mackay *et al* 2012). The lines are thus effectively homozygous at any given locus. However, because Drosophilids are highly panmictic (*i.e.* breed freely in the wild), the variation among the lines is proportional to, and reflective of, the substantial variation within the population from which they were derived, something that cannot be guaranteed in many populations (including humans). 81 genotyped

lines were selected from the DGRP collection based on three criteria: 1) including lines with the highest quality and depth of genome sequencing, 2) avoiding pairs of highly-related lines and 3) avoiding lines with unusual levels of residual heterozygosity.

6.1.2 CAGE Protocol (Ignacio Schor, Jacob Degner)

CAGE libraries were prepared according to the 5' CAGE protocol used in Hoskins *et al* (2011). Reads were demultiplexed and processed as per Takahashi *et al* 2012. The 26nt tags from the different DGRP lines were all mapped to the *Drosophila melanogaster* reference genome (dm3 assembly), using BWA version 0.6.1-r104, allowing for up to one mismatch (bwa aln parameter n=1). Mapped reads were filtered for those mapping to a single location in the genome and for overall mapping quality by filtering mapped reads for quality score > 10 (samtools view parameter –q 10). Finally, bam files were sorted and indexed for rapid access.

6.1.3 3' Tag-seq Protocol (Enrico Cavanno, Nils Koelling)

3' Tag-seq libraries were prepared according to the 3'-Tag-seq protocol used in Derti *et al* (2012). Ten samples were multiplexed and sequenced, and reads were demultiplexed allowing for one mismatch in the barcode, and mapped to personalized *Drosophila* genomes using BWA version v0.6.2-r126, with alignment parameters as described in section 'Identifying the location and expression of polyadenylation sites'.

Personalized genome files for each DGRP line were generated from the *Drosophila* reference genome (BDGP5) by replacing the reference genotypes with the alternate genotypes as annotated in the unfiltered, binned Freeze 2 DGRP variant annotation (freeze2.bins.vcf.gz, obtained from https://www.hgsc.bcm.edu/arthropods/*Drosophila* -genetic-reference-panel). For all 3'Tag-seq analysis, version 5.47 of the Flybase annotation was used.

6.1.4 Identifying the location and expression of pA sites (Nils Koelling)

Regions of polyadenylation, called pA sites, were defined in a two-step process: First, we identified the location of all pA sites measured by 3'-Tag-Seq, based on reads from all 254 samples using a method similar to that of Smibert *et al*[7]. Second, we determined how strongly each pA site was expressed in each individual sample.

To identify the location of pA sites genome-wide, we first found all reads which covered the 3' end of a transcript, followed by a stretch of non-genomic poly(A). For this, we mapped all 3'-Tag-Seq reads to personalized genomes using BWA version v0.6.2-r126[6], allowing for up to five mismatches, maximum 10 gap extension and a sequence quality threshold of 20.

The personalized genomes used for this step were generated as described in the section above, considering all annotated homozygous SNPs but not insertions/deletions. This was done to keep the coordinate system in sync between all lines, allowing us to merge the aligned reads from all samples and thus simplify the procedure outlined below. For the quantification of expression levels, we took both insertions/deletions and SNPs into account.

We then identified all unmappable reads that ended in at least five A nucleotides. From these reads, we trimmed off all trailing A nucleotides and then aligned them to the genome again. Any read that did not align in its original form but did align in its trimmed form with mapping quality >= 23 was considered a polyadenylated read. Across all 254 samples, we identified approximately 30 million of these reads. We then merged the poly(A) reads from all samples and calculated the coverage of poly(A) reads along the genome. Polyadenylation sites (pA sites) were defined as regions where the coverage was >15 overlapping reads, which were then expanded by 200bp upstream or up to the next upstream peak, whichever was shorter, to account for the expected length of the sequenced fragments. The 3'-Tag-Seq protocol uses a 200bp size restriction for the generated reads, and as expected, 92% of mapped reads fall within 200bp upstream of the leftmost cleavage event of a pA site.

The most likely annotated transcript for each pA site was identified by comparing

the pA site location to the Flybase annotation (v5.47) of all genes in non-heterochromatin regions of chromosomes X, 2, 3 and 4. First, we tried to find an annotated 3' end of an mRNA within 500bp of the pA site. If there was such an annotation, we associated the pA site with that transcript. If not, we looked for other possible annotations, assigning the pA site to the first annotation in the following order: ncRNA within 500bp, 5' end of mRNA within 500bp upstream, exon, intron, 3' end of mRNA within 2000bp upstream, 3' end of mRNA within 500bp on the antisense strand, 5' end of mRNA within 500bp upstream on the antisense strand, exon on the antisense strand, intron on the antisense strand.

Once the locations of pA sites were defined, we then aligned all original reads, trimmed to the minimum read length of 43bp, to fully personalized genomes using BWA with the same parameters as described above. For each pA site, we determined the height of the summit (the maximum number of overlapping reads within a 200bp window) for each sample using bedtools v2.16.2. We considered this height, scaled by the 90th percentile for each individual to normalize for library size, the raw expression level of the pA site. To obtain per-gene expression levels, the expression levels of all pA sites that could be assigned to a unique gene were summed.

6.1.5 3' Tag-seq: reducing mappability issues (Nils Koelling)

Sequence variation between individuals or lines can introduce variant dependent differences in the ability to correctly map short sequence reads back to their location in a reference genome (mappability). These genotype specific mapping biases can lead to artificial associations between genotype and any experimental measurement made with short read sequences. To address this potential bias, we used a four-step approach to reduce mappability issues as much as possible:

First, for each of the 80 DGRP genotypes we mapped all reads onto personalized genomes, as described in section 'processing of raw reads'. The advantage of this has already been demonstrated in a number of contexts.

A very stringent and effective method to address mappability biases is to identify the genomic locations for which different genotypes have different mappabilities

and remove these genomic locations from analysis. In line with this, we created a genome-wide map of the mappability of DGRP lines for each position in the genome and used this to remove QTL that were likely to be artificial associations. Raw sequencing reads from the DGRP genome sequencing project (75bp paired end Illumina reads) were used to generate all possible 26 bp short reads that could be obtained from each sample. These shortened reads were mapped to the reference genome (BDGP5) using BWA version v0.6.2-r126, allowing for up to five mismatches, maximum 10 gap extension and a sequence quality threshold of 20 and the number of reads starting at each position was counted for each line. We considered each position with at least one initiating read as mappable. We then removed any QTL where the mean mappability of the associated peak(s) was significantly different (unadjusted $p < 10e-6$) in a Wilcoxon rank sum test between lines with the homozygous major and the other genotypes.

Third, we determined the genome-wide heterozygosity across all lines used in our study by calculating the percentage of heterozygous genotype calls (made by DGRP) for each annotated variant. We removed every QTL that was associated with a region that contained a variant with heterozygous genotypes for more than 40% of the lines. Finally, for the gene-QTL, all QTL were removed where the significant variant was within our point of measurement (pA sites), as follows: We removed all SNPs within or close to (+/- 25bp) a pA site. As the pA sites are on average 200bp, this removed all variants within a ~250bp region around our point of measurement. This stringent 'in peak' filter created a very conservative set of eQTL. This last filter was not applied to the 3' isoform QTL (3iQTL) as it would remove the bulk of the biological regulation occurring in the 3' UTR.

6.1.6 Processing of 3' Tag-seq expression levels for QTL calling (Nils Koelling)

Before QTL calling, the data was processed to reduce potential confounders in a procedure similar to Degner *et al* 2012. This expression level normalization was applied to pA site expression levels as well as gene expression levels separately, and also to the data from each developmental time-point individually. First, all non-

expressed features were removed by filtering features for which we did not observe at least one read in at least 50% of the samples.

The expression data was then converted into z-scores by mean-centering and scaling the expression levels for each feature to unit variance. These scores were quantile-normalized within each individual line to follow a normal distribution. Finally, any remaining hidden structure in the data was removed using PEER (Stegle *et al* 2012), modeling effects of up to 10 hidden factors. The residuals estimated by PEER were used as our phenotypes for QTL testing.

6.2 Software packages used

Analyses and visualization in this thesis, unless otherwise stated, were carried out using custom scripts in R (v3.0.0), built upon primarily the GRanges (version 1.22.1,Lawrence *et al* 2013), ggplot2 (ver 1.0.1,Wickam 2009) and dplyr (version 0.4.2,Wickam 2014) packages.

6.2.1 Calling CAGE peaks

We found that the hierarchical clustering based algorithm used by Hoskins *et al* (2011) was unsuitable for our data because of its extremely high depth and the large range over which gene expression varies, which meant that many genes were tagged on a very high fraction of their bases, while others showed only a few tagged base pairs. We therefore needed a method that would account for the density of tags over a location, rather than simply clustering sites with above a certain threshold of CAGE tags.

Like Hoskins et al, we observed a large amount of noise distributed over the bodies of genes, particularly highly expressed genes. We also observed that this noise was not well reflected by existing RNA-seq datasets (e.g. Brown *et al* 2014), because of their relative sparseness, compared to our ultra-deep CAGE data. We instead chose to approximate such an RNAseq-based background by assuming that for each gene some fraction λ of its reads were distributed randomly across its length, and optimizing this fraction by the same iterative procedure used by Hoskins *et al* (2011). To estimate lambda for each gene, we first chose the value minimizing the total distance between predicted noise tag values and actual values over all base pairs. We then calculated the probability of each base pair having i's current tag values under the null, and excluded all values with P>1e-5. We then re estimated lambda for each gene, after excluding sites with the new set of putative noise sites and repeated the procedure. This process was iterated three times for each time point, and further iterations did not alter the estimated value of lambda.

With background noise due to highly expressed genes removed, we then derived peaks by first smoothing the data (taking the local mean over 150bp) and then selecting regions between local minima over 150bp. The window size 150bp was empirically selected as giving good separation between closely spaced CAGE peaks. The resulting peaks were then filtered to those having a read count of 50 or more across all libraries.

We then wished to link our peaks to known transcript models. We reasoned CAGE peaks upstream of known genes were likely to be un-annotated peaks, while peaks in the body of known genes were more likely to be due to re-capping or experimental artifacts. We therefore defined a zone of interest for each gene extending 250bp downstream of annotated TSS, and up to 2kb upstream (while not including other genes on the same strand). Peaks overlapping these regions and on the correct strand were included in our 'main' set, and were linked to the corresponding gene. Where a peak was in this region for multiple genes, we linked it to the gene with the closest TSS. Also included were those extragenic peaks overlapping a DNase hypersensitive region (Thomas *et al* 2011) that may correspond to promoters of annotated genes or eRNAs. Peaks excluded from the 'main' peak set but overlapping gene bodies in the same strand were annotated as 'internal' peaks. Visualizations of CAGE signal over gene bodies and peaks were created using the R Sushi package. For all CAGE analysis, the UCSC dm3 gene annotation was used.

6.2.2 Gene ontology enrichment/depletion analysis

To determine which genes were statistically enriched among genes with many 'main' peaks and among genes with many up-regulated clusters, we used a point-biserial correlation test. As input for the former, we used the residuals from a negative binomial model incorporating peak number as the response variable and the genes' expression and length as covariates. As input for the latter, we used the residuals from a linear model including gene expression and length as covariates. Individual genes were mapped to pathway and GOSlim ontology terms using classifications provided by the PANTHER Classification System (version 8;

http://www.genome.org/cgi/content/full/13/9/2129). Because these ontological categories are extensively nested, and thus not independent, standard methods for multiple testing correction cannot be easily applied. We thus chose to keep all categories with nominally significant p-values ($p<0.05$), and clustered them graphically by term similarity to better assess the number of independently significant categories in each test group.

6.2.3 Shape Index

We calculate the entropy of the distribution of TSS at a promoter of n bp by treating it as a discrete distribution with n possible outcomes, and using the frequency of CAGE tags at position x[i] to estimate P(x[i]):

$$-\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i),$$

We follow the convention of Hoskins *et al* in subtracting the entropy from the constant 2 to yield a 'Shape Index':

$$SI = 2 + \sum_{i}^{L} p_i \log_2 p_i \,,$$

6.2.4 CAGE Peaks – differential expression and promoter usage analysis

Differential expression analysis over timepoints was carried out using DESeq 2 (Love *et al* 2014), by contrasting between the first time point (2-4 hours) and the last time point (10-12 hours). We treated the CAGE libraries for individual lines as biological replicates. We counted peaks as having significant differential expression as long as the absolute value of their log2 fold change was 1.5 or higher.

6.2.5 Changes in 3′ UTR length during development

To examine 3′ UTR length we considered only those pA sites with a median 3′-Tag-Seq count > 0 in at least one time-point (i.e. pA sites with >= 1 tag in half of the 80 lines). We attributed each of the pA sites to genes as described above. We additionally extended out to a further 10kb upstream of unattributed genes and assigned pA sites to the first gene found in the correct orientation, thus allowing pA sites to be located up to 10kb from their putative parent gene.

To examine 3′ UTR length in general, we calculated the unspliced 3′ UTR length for all pA sites, using the following procedure: Only pA sites that were 3′ of an annotated stop codon from their parent gene were taken. The closest annotated stop codon for each pA site was used, to get a putative 'unspliced UTR length'. We then took the maximum such length for each gene, to get a 'maximum unspliced UTR length'. The small minority of cases where this distance was > 10kb were excluded, as they were generally associated with early stop codons at the start of large genes and we reasoned that these distances more likely reflect unknown isoforms, rather than actual UTRs.

To obtain a stringent set of genes whose 3′ UTR length changes during development (shown in Fig. 5b), differential usage of pA sites was taken into account. This avoids cases where there is a distant minor pA site with low expression, which would skew the results for a gene with generally short UTRs. To calculate the change in UTR length for each gene, we first estimated the mean length of the 3′ UTRs expressed by a gene at each time-point by taking the length of each of its unspliced UTRs and weighting them by their quantified expression. The fold change was then measured in mean unspliced UTR length between the earliest (2-4 hr) and latest (10-12 hr) time-point. To avoid this measure being affected by high variance of weakly expressed pA sites, only pA sites that were expressed in half or more of the lines, at all time-points, were included (Table S14). GO term and expression enrichment/depletion analysis is described above.

6.2.6 Transcription factor motifs

TF position weight matrices (PWMs) were gathered from the following sources: a) our own lab's data[31,39,40]; b) the modENCODE consortium[42,45]; c) Berkeley TF ChIP-Seq data[43]; d) Berkeley *Drosophila* Transcription Network Project http://bdtnp.lbl.gov/Fly-Net/selex.jsp?w=summary; e) Flyfactor/Flyreg databases, http://pgfe.umassmed.edu/ffs/, http://bergmanlab.ls.manchester.ac.uk/flyreg/; f) the Jaspar database, http://jaspar.genereg.net/. This resulted in an extensive collection of 1,025 *Drosophila* based PWMs, of which 337 corresponded to 73 unique TFs with matching ChIP datasets (Table S8). These were used for the rest of the analysis. As no count information was available for the modENCODE PWMs, we followed the authors procedure for generating PWMs from frequency matrices, and used a very small pseudocount, but with genomic base frequencies, rather than frequencies of 0.25. We first multiply the elements of each frequency matrix by 1000, to yield count matrices, and then add a pseudocount of one before dividing by the total count per position to yield adjusted frequency matrices. The final formula we then use for each position is:

$$M_{i,j} = \ln(F_{i,j} / B_j)$$

where $M_{i,j}$ is the value for base j, position i in the PWM , $M_{i,j}$ is the value in the adjusted frequency matrix, and B_j is the frequency of base j in the genome (i.e. 18.2 for G/C, and 32.2 for A/T).

6.2.7 Promoter-associated motifs

Promoter-associated motifs were gathered from the following sources. PWMs for the eight Ohler motifs were taken from the supplemental material of Ni *et al* 2008. IUPAC motifs were used for the 15 motifs identified by Fitzgerald *et al* 2006. The motifs identified by Down *et al* 2009 were downloaded from the 'Tiffin' database. Since these motifs were available only in the form of frequency matrices rather than counts, we simply multiplied the frequency matrices by 100 so that they could be processed along with the other motifs.

6.2.8 RBP motifs

We took information on protein binding motifs from Ray *et al* 2013. We followed their procedure in using the protein binding microarray data directly, and counting as a match only 'strong' n-mers with a Z-score of 0.45 or more.

6.2.9 De novo motif discovery

We used the Meme Suite (Bailey *et al* 2009), for all de novo motif discoveries. Meme-chip was called on target regions, using a maximum motif size of 15bp and an E-value cutoff of $5x10^{-5}$. Meme-chip uses both the generative Meme algorithm to find enriched PWMs, and the enumerative DREME algorithm to find enriched IUPAC words within a set of target sequences. The Meme suite also includes Centrimo, for assessing positional enrichment of a motif within a set of sequences, and TomTom, for finding matches to similar motifs within a database. For 3' Tag-seq peaks we ran meme-chip simultaneously on the 300bp region centered on the 3' end of each pA site. For promoter-associated motifs we ran meme-chip on CAGE peaks, centering on the most highly tagged site within each one and extending outwards +-250bp. For promoter-associated motifs, separate runs of meme-chip were carried out on broad CAGE peaks, narrow CAGE peaks, and their union, as well as runs contrasting broad against narrow, and vice versa. We ran Centrimo on all discovered promoter-associated motifs, plus the three sets of promoter motifs derived from the literature, to yield central enrichment scores for each one within the set of all TSS. We found that TomTom was over conservative when scoring motif similarity and failed to match many variants of motifs with identical positional enrichments. Motifs were assessed for similarity using the column wise correlation score described in Pietrokowski 1996. To this we added 0.1 where motifs had overlapping positional enrichments, to yield a similarity score. From this score we constructed a similarity matrix and performed complete hierarchical clustering, cutting the tree at a distance of 0.6 to yield clusters of similar motifs.

6.2.10 Scanning for motifs

To assess prevalence of motifs, we scanned the genome using Patser (Stormo *et al* 2000) for PWMs, and the Biostrings R package for IUPAC motifs. We counted motifs only where they were present within the zone of positional enrichment defined by Centrimo. Motifs were assessed for strandedness by testing their presence on each strand within this zone, and those with a significant bias towards the reverse strand (as assessed by a binomial test) were reversed, so that all motifs were present in the correct orientation relative to the pA site or promoter.

6.2.11 ROC analysis

In order to (a) assess the quality of the PWM motif and its relevant ChIP dataset for each TF and (b) determine a relevant cut-off for each TF, ROC analysis was used to derive sensitivity curves and AUCs for each factor. True positives were defined as the significant ChIP peaks for that TF, while false positives were defined as DNaseI hypersensitive regions (Thomas *et al* 2011) not bound by that TF, as described below. Regions of exactly 400bp were used, with the provided summit of the ChIP peak being used, or its centre where no summit was available. ModENCODE defined hot-regions and blacklisted regions were excluded, in addition to 3'UTRs and heterochromatin, as previously described (Negre *et al* 2009). DNase regions that overlapped a TF motif, another motif of which closely matched the motif, were also excluded from negative sets in order to allow accurate assessment of enrichments in e.g. motifs for factors with similar binding preferences to other factors. We proceeded to use the motif cut-off that gave 50% sensitivity for our analysis. Experimenting with other thresholds, such as those based on a fixed p-value or false discovery rate, did not substantially improve enrichment scores.

6.2.12 Enrichment score Analysis

Our enrichment score analysis was based on the same measured used by Negre *et al* (2009). We first generated shuffled motifs for each of our PWMs by permuting the

columns of the PWM. Using the same HOT regions and background regions from our ROC analysis, we counted motif matches using the 50% sensitivity threshold derived from the ROC analysis inside and outside matching ChIP peaks, for both shuffled and non shuffled motifs. Using 95% binomial confidence intervals for these two fractions, we then derived the most conservative possible value for their ratio, and defined this as the enrichment score. Motifs with an enrichment score of 1.1 or greater, and an AUC of 0.6 or more, were used in further analyses.

6.2.13 INSIGHT analysis

INSIGHT analysis was carried out using scripts provided by Gronau *et al* 2013. The scripts require input on both intra species variation – for which we used the complete genotypes of the DGRP lines (Mackay *et al* 2013), and the 12 sequenced *Drosophila* genomes (Clark *et al* 2007). We modified the scripts to use nearby 4d degenerate sites rather than flanking intergenic regions as neutral proxies, as the high density of the *D. melanogaster* genome means that most of these regions are not evolving neutrally. Rho-enrichment scores were calculated by taking the 95% confidence intervals given by INSIGHT, and using them to derive the most conservative possible ratio between the rho values for shuffled and non-shuffled motifs, as with the enrichment scores. A similar procedure was used to assess enrichment for adaptive evolution.

6.3 Enhancer transcription

6.3.1 Enhancer transcription analysis in S2 cells

GROseq data for S2 cells was taken from Core *et al* 2012, and GROseq data for IMR90 cells was taken from Core *et al* 2008. Short capped nuclear RNA data was taken from Nachaev *et al* 2010. PROcap data fro S2 cells was taken from Kwak *et al* 2013. CAGE data for S2 cells was taken from Cherbas *et al* 2011. DHS and STARR-seq data for S2 cells were taken from Arnold *et al* 2013.
Directionality index was calculated as (Max / (Min+Max)) where Max is the strand with the most reads for the element, and Min is the opposite strand.

6.3.2 Enhancer transcription analysis in Whole embryo

DHS for whole embryo analysis were called using data from Thomas *et al* 2011. Transgenic enhancers used were a database consisting of transgenic enhancers gathered from the literature with known tissue specific activities (Zinzen *et al* 2009, Kvon *et al* 2014). PROcap data was prepared according to the protocol used by Kwak *et al* 2013, and processed similarly to the CAGE data (credit - Ignacio Schor).

6.3.3 In vivo expression assays (Olga Mikhalylichenko)

Enhancers were selected from the 8008 enhancers in Zinzen *et al* 2009, and filtered with ChIP data for K27ac and PolII using data from the modEncode project (Kharchenko *et al* 2011). For transgenic expression assays, selected enhancers were PCR amplified from wild type *Drosophila* genomic DNA and subcloned in pCRII-TOPO vector. Modified PhiC31 integrase targeting vector containing an attB recombination siteand and a loxP site was digested with one of the following combinations of restriction enzymes: 1) XbaI + Kpn I (to generate plasmids containing hsp70 minimal promoter) or 2) XbaI + PstI (to generate plasmids without minimal promoter). Each enhancer was cloned in two vector versions (with and without minimal promoter). Resulting strains (17 in total, including promoter-only control vector) were

generated by injecting eggs from J27 *Drosophila* line under standard procedures. All constructs were injected according to standard methods into the J27 line. Stably integrated transgenic lines were balanced, homozygosed and used for embryo collections, and subjected to *in situ* hybridization to examine *lacZ* expression. Transgene expression patterns were detected by fluorescent *in situ* hybridization (FISH) (Kosman *et al* (2004)) using an antisense fluorescein-labeled RNA probe directed against LacZ transcript. Embryos were counterstained for *Mef2* RNA with a digoxygenin (DIG)-labeled probe.

6.4 CAGE QTL calling (credit - Jacob Degner)

CAGE QTL windows were called on 1000bp windows, independently of gene annotation. First, all individual CAGE libraries were combined into a single bam file. From this combined data, we used a greedy search algorithm to select a minimum number of ~1000bp windows, which together contained more than 99% of the total, CAGE reads. This greedy algorithm behaved as follows

1) Tally number of reads originating from each base in the genome

2) Select the single base with the highest read count

3) Identify the 1024 base pair region centered on this single base as the next phenotype window, replace count at all bases contained in this window with 0 in the tally made in step 1. Record the total fraction of all reads contained in phenotype windows

4) Repeat 2-3 until total fraction of reads contained in all chosen phenotype windows was greater than 99%

6.4.1 Generation of phenotypes for CAGE-QTL analysis (credit - Jacob Degner)

For each of the phenotype windows selected in the previous step, we summarized the CAGE signal in two different ways (PC-based and Mean based).

Mean based approach

1) For every base contained in the phenotype window (excluding those determined to be unmappable as described above), count the number of CAGE reads starting at that position. Do this separately for each CAGE library and where multiple CAGE libraries exist for the same individual line at the same timepoint, add together the individual matrices.

2) Normalize these count matrices by the total sequencing depth for that individual/timpoint combination

3) Average normalized read counts across the bases contained within this window and project quantiles of the distribution across individual/timpoint

combinations onto the quantiles of a standard normal distribution (to adhere strictly to the assumptions our statistical model)

This procedure results in a single measurement for each timepoint/individual combination and these measures were used to test for associations in the multi-phenotype linear mixed model framework described later.

Verifying genotype labels from raw 3' tag-seq reads

Prior to merging different libraries for association mapping, the sequence reads for each independent library preparation were checked against the published Freeze2 genotypes given by the DGRP project to verify that the identity of the sample (Mackay *et al* 2012). For each tag-seq library, we extracted the base-call at every position within all reads overlapping SNP positions. Given these base-calls in each library, for each of the 210 DGRP lines, we calculated a summary statistic describing the fraction of tag-seq reads that were consistent with the genotype of each line. The distribution of this statistic was clearly bimodal with one mode representing mis-matched individual to library comparisons and the other mode representing matched individual to library comparisons. Several sample swaps were identified using this procedure, but in the end, only libraries that could be confidently identifiable as a different DGRP line and thus were re-labeled with the corrected DGRP line label.

6.4.2 CAGE processing for tssQTL - Creating a universal mappability map for the DGRP (credit - Jacob Degner)

Sequence variation can induce variant-dependent differences between individuals or lines in one's ability to correctly map short sequence reads back to their location in a reference genome. These sequence-specific biases in mapping can lead to artificial associations between genotype and any functional genomics measure made with short read sequences (Degner et al. 2009). The most effective method for addressing these biases is to identify the genomic locations for which different genotypes have different mappabilities and remove these genomic locations from analysis (Degner et al. 2009, Degner at al 2012). To address this potential bias, we created a genome-

wide map of the mappability of each DGRP line for each position in the genome and used this to pre-filter regions that are likely to introduce artificial associations. Raw sequencing reads from the DGRP genome-sequencing project (75bp paired end Illumina reads) were used to generate all possible 26 bp short reads that could be obtained from each line. These shortened reads were mapped to dm3 and the number of reads starting at each position was counted for each line. From these counts, a universal mappability map was created defining the genomic positions for which each DGRP line contained at least one read that mapped to that location. Finally, a summary of this map was created such that there were mapped reads represented at that position for all DGRP lines with modest sequencing depth (zero read counts were ignored if they only occurred for a single DGRP line and that DGRP line had an average sequencing depth < 10X)

6.4.3 PC-based approach (credit - Jacob Degner)

1) For every base contained in the phenotype window, count the number of CAGE reads starting at that position. Do this separately for each CAGE library and where multiple CAGE libraries exist for the same individual line at the same time point, add together the individual matrices

2) Normalize these count matrices by the total sequencing depth for that individual/ time point combination

3) To limit the impact of extreme values on determining the direction of the first principal components, square root transform the normalized base-level read counts

4) Construct a total matrix for a single phenotype window with individual/ time point combinations as rows and bases as columns

5) Decompose this matrix into its principal component representations

6) Extract top three principal components and as before, convert the distribution of these measures to the corresponding quantiles of a standard normal distribution.

This procedure results in three measurements for each individual/ time point combination (or nine total phenotypes per *Drosophila* line). These were also tested for association with genotypes using a multi-phenotype modeling framework described below.

6.4.4 Estimating single base effect sizes of significant QTL with waveQTL (credit - Jacob Degner)

While the PC-based approach proved to be powerful and scalable to the huge number of individual SNP X phenotype evaluations performed here, it had several drawbacks. First, only the first three PCs were chosen for testing which while capturing the dimensions of highest variance, discard many more dimensions represented in the full data than they retain. Second, we expect the biology of transcription start site choice to at least somewhat spatially dependent (Indeed, CAGE signal is known to have substantial autocorrelation between nearby bases). The decomposition of signal into PC representations is completely ignorant of the spatial relationship between the bases (i.e., PC values used as phenotypes here are identical if we scramble the order of bases before performing the decomposition). These two limitations are effectively addressed in a recent method called WaveQTL (Shim *et al* 2015). Instead of decomposing the matrix of CAGE signal into PC-based representation, the signal is decomposed into projections onto wavelet space and using the full representation of signal in this space, posterior distributions on effect size are estimated for each base. As wavelets capture successively larger chunks of the base-pair space, estimation of effect sizes in this space allows spatially restricted sections of the phenotype window to share the same effect size when effects are distributed on the lower-order (larger segment) wavelet coefficients. This procedure results in estimations of effect size both in base-pair space and wavelet space and we use both of these below (note that estimation is done in wavelet space which can be subsequently transformed to base-pair space.)

6.4.5 Classifying QTL according to their pattern of wavelet and single-base effect size (credit - Jacob Degner)

We noticed from exploring the raw data and predicted effect sizes of QTL that there was a range of QTL whose biological interpretation seemed very different. For example, we noticed that some QTL, especially those uniquely identified by the PC-based approach, did not seem to affect overall transcription, but simply changed the position of the bases used as transcription start sites. We began referring to these QLT as 'Redistribution' type QTL and devised a set of rules for distinguishing them from the opposite 'Directional' QTLs. As effect size estimation with WaveQTL is performed in wavelet space, there is a direct estimate of the effect on the highest spatial subdivision of the data that corresponds to an effect size estimate on the mean level of CAGE signal in the whole phenotype widow. Thus, to classify a QTL as 'Redistribution' type, we required the overall evidence to be in favor of no association with the mean CAGE level (log Bayes Factor < 0 for evidence of association with the lowest wavelet coefficient.) In addition, we found that requiring the log10 bayes factor of at least one wavelet coefficient to be > 1 reduced the number of very difficult to interpret QTL. On the other extreme, we classified a set of purely 'Directional' QTL where log10 BF > 1 and all the bases with a significant effect size were in the same direction (significance at the base-pair level corresponds to a 95% credible interval on base-level effect size that does not overlap zero).

6.5 3' Tag-Seq QTL calling (credit – Nils Koelling)

For each gene (gene-QTL) and pA site (3' isoform QTL – 3iQTL), we tested all annotated biallelic variants in a proximal (putative *cis*) window spanning at least 100kb (50kb upstream, the entire gene locus and 50kb downstream) around the region and with a minor allele frequency (MAF) of at least 5% for association with the observed expression levels. Testing was performed using a mixed linear model framework that accounts for developmental stage and population structure, by comparing the likelihoods of the following linear mixed models Lippert *et al* 2014.

Briefly, the normalized expression data from all three developmental stages were modeled by the sum of a random effect that accounts for relatedness and a noise component, while simultaneously accounting for genetic and non-genetic trait-to-trait correlations. Based on this null model, genetic associations were tested using fixed effect tests, considering the following testing designs: (1) Common effect test: a single fixed effect (one degree of freedom) with a shared effect size across all developmental stages was tested against the null model without fixed effect covariates. (2) Specific effect test (for each of the three developmental stages): A specific fixed effect covariate with an effect only in one of the three stages was tested, where the common effect test (1) serves as null model.

The common effect test and the three specific effect tests (for each developmental stage) were applied to each variant/region, resulting in p-values for each of the four tests. For each of these QTL types, we separately adjusted for multiple testing per *cis* window using 10,000 permutation experiments, estimating an empirical p-value that corresponds to the hypothesis that at least one variant in the respective *cis* window is associated. For each region, we chose the variant with the lowest p-value of the given specificity as the lead variant.

To account for multiple testing, we then adjusted the empirical p-values of the lead variants across the set of all lead variants from the four different QTL types combined using Benjamini & Hochberg's method to obtain a global, experiment-wide false discovery rate (FDR).

For all downstream analyses, we only considered QTL at a significance level of FDR < 10%. To account for the possibility that the lead QTL might not be the causal variant, we additionally generated an extended QTL set containing each region's lead QTL as well as every other significantly associated QTL with a p-value within one order of magnitude of the top p-value for that region, as long as its empirical p-value was still above 1%. In the following, we denote this extended set as QTL clouds.

Effect sizes of common effects were directly obtained from the common effect test across all three developmental stages. For stage-specific QTL, we retested the common effect model using only the data from a single developmental stage and

considered the effect size from the variant effect term in that single-stage model as the effect size.

This procedure resulted in four sets of QTL for each region type (for both **gene-QTL and 3iQTL**) – one set of QTL with a common effect and three sets of QTL with a stage-specific effect, one for each developmental stage. We flagged all stage-specific QTL that also had a common effect as secondary effects and did not consider them when calculating the total number of stage-specific QTL. Furthermore, we classified stage-specific QTL based on their single-stage effect sizes, which were obtained by re-testing the common effect model a single stage at a time as described above: QTL with more than one specific effect were classified as complex specific QTL. QTL for which the largest single-stage effect size was estimated in their associated stage were classified as single-stage QTL. All other QTL were classified as weak stage-specific QTL, and are likely the result of being present at two stages, and specifically absent at the stage in which they were called.

3iQTL were associated to the gene to which the pA site was assigned. If genomic features of multiple genes were equally close, one of the genes was chosen at random. We applied the set of mappability filters to each set of gene-QTL and 3iQTL, as described in section “Reducing mappability issues”. In addition, we removed all QTL clouds that contained QTL further than 10kb apart or numbering more than 10 in order to remove regions of inferred long potential LD where we were unlikely to be able to identify the causal variant. 1,755/4,125 (43%) of gene-QTL and 6,995/12,114 (58%) of 3iQTL passed both the mappability and the cloud filters, and are provided in Tables S5 and S11, respectively. In these final sets, each gene contained on average 2.2 QTL in its QTL cloud, while each 3iQTL cloud contained 2.1 QTL on average.

To identify QTL that specifically affect mean UTR-length (**UTR-QTL**), we first estimated the UTR-length phenotypes by taking the mean length of a gene’s unspliced UTR, weighted by its pA sites’ expression. We performed this procedure individually for each gene, genotype and time-point. For each gene and developmental stage we tested for association between mean UTR-length and all variants in a window of 50kb upstream and downstream of the gene using a linear

mixed model analogous to the approach used for gene level QTL. Correction for multiple testing per *cis* window was performed using the LIMIX implementation of Q-values (Lippert *et al* 2014). This led to the identification of 3,821 genes with a mean UTR-length QTL at a nominal gene-level FDR of 0.01. To remove cases where changes in mean UTR length are fully explained by changes in expression, we only considered genes for which we still identified a significant association ($p<0.01$) between mean UTR-length and the lead variant after introducing gene expression (PEER residuals of gene expression levels, quantile-normalized to a normal distribution) in the model as a fixed effect covariate. This filter led to a stringent set of 2,005 genes with a UTR-QTL. For genes with a UTR-QTL at multiple developmental stages, only the most significant association was considered for further analysis, defined as the one with the lowest gene-level FDR. Similar to the gene-QTL analysis described above, to remove regions of inferred long potential LD, all genes for which more than ten variants had a p-value within one order of magnitude from the lead variant, were removed. UTR-QTL were further filtered for mappability of pA sites associated with the gene (maximum heterozygosity 40%, no correlation between genotype and mappability < 1e-6, as described for the gene-QTL). For this high-confident set of 764 UTR-QTL, the change in mean raw UTR length between the major and minor allele was calculated as the difference between the median mean UTR length among all individuals with the minor allele and then median mean UTR length among all individuals with the major allele. To focus on UTR-QTL that caused a substantial change in base-pair length of the 3' UTR, we only considered the 311 cases where the absolute value of this difference was greater than 25bp between the major and minor allele for further analysis.

6.6 Motif change analysis

I wrote a custom pipeline that allows for the assessment of changes to PWM scores or IUPAC motif matches between genotypes, without the need for the construction of alternative genomes. The pipelines takes advantage of the fact that for a motif of size N bp, two mutations, if they are spaced greater than N bp apart, cannot affect the same motif match and can therefore be analyzed independently. In effect, local alternative genomes can be constructed for each haplotype, and the results from each haplotype then mapped back to the reference genome. This operation is somewhat trivial for SNPs, but becomes nontrivial for indels, and particularly combinations of indels. We found that in general the number of haplotypes with mutations affecting the same motif was very low in our data, and so we excluded these effects from subsequent analysis, treating mutations individually.

When analyzing changes in motif scores for TF motifs, we counted changes of 1 nat or greater to the motif score (Patser motif scores are in base e), and (for the 3' Tag-seq analysis) counted changes of 3 nats or more as 'strong changes'. Destructions of motifs were counted only where they overlapped a peak for a matching ChIP dataset, while destructions of motifs were counted if they overlapped a DNAse sensitivity peak.

When analyzing IUPAC or CisBP motifs for 3' Tag-seq motifs, we required perfect matches, and when analyzing IUPAC promoter-associated motifs, we allowed a mismatch of 1, since the number of tssQTL was smaller than the number of 3'iQTL. Care must be taken when defining motif changes as 'destruction' and 'creation' since the result will depend on the genotype used as reference. Because the LIMIX framework defines eQTL effects sizes relative to the major allele, rather than the reference we defined our motif destructions and creations in the same way.

6.7 QTL-feature Enrichment: Logistic Regression Framework

In order to assess the enrichment of functional categories for eQTL, we used a logistic regression framework resembling that of Brown *et al* 2013. Ours differed from theirs in several respects. While Brown *et al* used feature overlap as their response variable, ours was of the form:

$$P(\mathrm{Eqtl}) \sim \mathrm{Logit}^{-1}(\beta_0 + \mathrm{Feature} + \mathrm{MAF} + \mathrm{Log}_{10}\mathrm{Expr})$$

Equation 3: Where Eqtl is a binary variable denoting whether a given Variant-Gene Pair are significantly associated, Feature is a binary variable denoting the variant's overlap with the feature, MAF is the variant's minor allele frequency, and Expr is the total number of 3' -Tag-Seq reads attributed to this gene.

Therefore, while Brown *et al* asked about the probability of finding a feature given the presence of an eQTL, our regression is informative about the probability of finding an eQTL given a feature. This latter probability is more complicated to estimate, but is more relevant, in that typically the functional relevance of a SNP is the variable of interest, rather than its inclusion in a feature, which is typically known. As input data, our framework uses a matrix where each row corresponds to a tested variant-region pair, where the region is either a Gene, a pA site, or a CAGE window. Regressions were calculated using the R package speedglm. We included an additional binary variable denoting presence under the pA site for 3iQTL enrichments, to control for mapping biases. Our framework thus accounts for the differential size and polymorphism of the different features compared, as well as differential power due to the expression of each gene (taken as the sum across all time-points) and the variant's minor allele frequency (MAF). Since both high expression and high MAF increase our ability to detect QTL, and both are expected to vary across different features, controlling for these factors is necessary to avoid biasing the results of the regression. For 3iQTL feature enrichments (Fig 4a,d), the

logistical regression framework also includes, as control variable the expression of each pA site and MAF. Here, all genetic variants were included as we reasoned that (a) much of the biological signal we were interested in is located close to pA sites and (b) potential genotypic variants causing mapping issues should not have a biased distribution towards particular sequence features.

6.8 Global feature enrichments

For gene-QTL feature enrichments (Fig 2h, 3b), as our high-confidence QTL set (test set) was generated by excluding QTL at the point of measurement, (~250bp near the 3′ -Tag-Seq peak, as described above), we also excluded genetic variants within these distances (extending by +/- 25bp) from all pA sites (background set). Broad feature enrichments (Fig. 4a) were carried out in an identical fashion to the gene-QTL, using variant/pA site pairs rather than variant/gene pairs. Gene-feature enrichments for tssQTL were carried out by the same method. UTRs were excluded from the exon category, and random control regions were included to verify that the regression framework gave unbiased results.

CRM feature enrichments for 3'-Tag-seq eQTL were calculated by using a logistic regression framework on tested gene-variant pairs but separating proximal variants from distal variants, where distal variants were those 1kb or more from a TSS for the eQTLs target gene. DNase peaks used were a set of peak calls made using HOTSPOT on data taken from Thomas *et al* 2011 merged over all timepoints overlapping with our expression data. while ChIP peaks used were those defined as High Quality (see results chapter 2). Transcription factor binding sites were defined as matches to the high quality TF PWM set, which overlapped a relevant High Quality ChIP dataset, at the 50% sensitivity threshold defined during our PWM screening pipeline. We expanded our for 25bp in either direction, reasoning that some of our eQTL might be markers for lower frequency disruptions to eQTL that failed to pass the frequency threshold for eQTL calling.

For RNA-binding motif enrichments in 3iQTL, (Fig. 4d) we considered only those variants within the 300bp window centered on the 3′ end of our pA sites.

For tssQTL motif enrichments, only promoter-associated motifs with positional enrichment were used, and motifs were grouped together according to the location of their maximum positional enrichment relative to the TSS.

6.9 3i QTL plots (credit - Nils Koelling)

3i QTL plots (e.g. Fig. 2d, 4f,h 5f,h) were generated using custom scripts in R (v3.0.0; http://www.R-project.org/), built upon the Rsamtools (v1.12.3; http://bioconductor.org/packages/release/bioc/html/Rsamtools.html), ggplot2 (v0.9.3.1) and ggbio (version 1.8.5), packages. Median 3' Tag-Seq coverage plots show library-size adjusted read coverage on the sense strand within the 3i peak regions associated with the gene only. Median RNA-seq coverage plots show library-size adjusted read coverage on the sense strand within the entire annotated gene body. Heatmaps (grey) show library-size adjusted read coverage on the sense strand in the entire region, scaled from 0 (white) to the maximum coverage observed in the region (black). Individuals with a heterozygous genotype call at the lead variant were omitted from all QTL plots. The p-values given above surrounding genes in QTL plots show the unadjusted p-value from the common-effect (or specific-effect, where applicable) test between the lead variant and expression of the surrounding gene. For 3iQTL, only the lowest p-value among all 3i peaks associated with the surrounding gene is shown. For UTR-QTL, the p-value from the association test between the lead variant and the mean UTR length at the indicated developmental stage is shown. A dash indicates that the surrounding gene was not tested with the lead variant, either because it was not expressed or because it was too far away. Values below 10^{-3} are shown in orders of magnitude.

6.10 tssQTL reporter validations (Ignacio Schor)

In order to allow for expression measurement of promoter strength and noise, we opted to measure the expression of a fast folding GFP variant in thousands of cells

using flowcytometry, for each construct. We transfected constructs into *Drosophila* S2, with GFP under the 3rd generation Tet-One™ Inducible Expression System (Clontech). We also used an mCherry protein expressed from the same plasmid under the actin5C gene promoter as an internal control.
Flow cytometry raw data was processed using FlowJo v.10.0.8 software. After selection of single live cells, we apply a threshold of 500 on the mCherry on mCherry value to exclude untransfected cells. Values for all positive cells were then exported in csv files, and further analysis was performed in R. Expression values per cell are calculated as the sfGFP/mCherry ratio for the selected cells. When log10-transformed, these resemble a normal-like distribution, although different promoters show different deviations from normality.

6.11 3'Tag-seq reporter validations (Enrico Cavanno)

The following regions were amplified and cloned into the pGL3-Hsp70 vector, or (Fig 4.10) downstream of the luciferase CDS. Site directed mutagenesis was performed using the QuikChange Lightning Multi Site-Directed Mutagenesis Kit (Agilent) accordingly to manufacture's instructions. The full length coding sequence of *slp1* was amplified using primers listed below and cloned into the pAc5.1 vector.

Luciferase assays were performed as described in Best *et al* 2014 using the dual-luciferase reporter assay system (Promega) according to the manufacture's instructions. Briefly, S2 cells were transiently transfected with Cellfectin (Invitrogen) to introduce plasmids carrying (i) the enhancer sequence linked to the hsp70 minimal promoter and luciferase (pGL3-Hsp70) (ii) a Renilla construct for normalization, (iii) When the TF was not expressed in S2 cells (i.e. *slp1*), the full-length TF under constitutive control (pAc5.1 vector) was co-transfected. In the latter case, the total amount of transfected DNA was kept constant by adjusting with an empty pAc5.1 vector. Levels of Luciferase and Renilla were measured 48 hours after

transfection with a PerkinElmer 1420 Luminescence Counter. For each transfection, three biological replicates were performed, each done in triplicate.

6.12 Relationship between tssQTL and Promoter Shape

Shape scores for CAGE windows – which were used for the tssQTL calls, were calculated via the same method as for CAGE peaks. We found that a small population of CAGE windows had low expression values, shape indices of -4 or less, and almost invariably overlapped internal CAGE peaks. We excluded these windows from the analysis. Logistic regressions of QTL likelihood vs shape were carried out in R, with the presence of a tssQTL (or subtype thereof) for a window as the response variable, and its log10(Tag Count)and shape index as predictors. The independence of Shape Index from expression was verified by a likelihood ratio test comparing a model containing shape index and expression to a model containing only expression. CAGE peaks were assigned motifs using the eight PWMs from Ni $et\ al$ 2010, with a permissive Patser p-value threshold of p=0.001. Motif hits were counted only where they overlapped the zone of positional enrichment for the motif defined by Centrimo. Proportion of functional sites (rho) and rate of adaptive substitutions for CAGE windows was calculated using INSIGHT (see 4.14). Windows were grouped by shape index into ten bins, and 4d degenerate sites were used as neutral control regions.

Bibliography

- A G Clark, L.W., 1997. Epistasis in Measured Genotypes: *Drosophila* P-Element Insertions. *Genetics*, 147(1), p.157.
- Alan K Kutach, J.T.K., 2000. The Downstream Promoter Element DPE Appears To Be as Widely Used as the TATA Box in *Drosophila* Core Promoters. *Molecular and Cellular Biology*, 20(13), p.4754.
- Andersson, R. et al., 2014. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493), pp.455–461.
- Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in *Drosophila* . *Nature*, 437(7062), pp.1149–1152.
- Arbiza, L. et al., 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nature Genetics*, 45(7), pp.723–729.
- Arnold, C.D. et al., 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, 339(6123), pp.1074–1077.
- Auer, P.L. & Lettre, G., 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome medicine*, 7(1), p.1.
- Austenaa, L.M.I. et al., 2015. Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. *Molecular Cell*, 60(3), pp.460–474.
- Babbitt, C.C. et al., 2010. Multiple Functional Variants in cis Modulate PDYN Expression. *Molecular Biology and Evolution*, 27(2), pp.465–479.
- Bailey, T.L. et al., 2009. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2), pp.W202–W208.
- Banerji, J., Rusconi, S. & Schaffner, W., 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2), pp.299–308.
- Best, A. et al., 2014. Tra2 protein biology and mechanisms of splicing control. *Biochemical Society Transactions*, 42(4), pp.1152–1158.
- Bin Z He et al., 2011. Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with *Drosophila* Cis-Regulatory Modules. *PLOS Genet*, 7(4), p.e1002053.
- Bloom, J.S. et al., 2000. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. 6 SP -, pp.–.
- Brawand, D. et al., 2011. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369), pp.343–348.
- Brown, J.B. et al., 2000. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, advance online publication SP - EP -, pp.–.
- Carninci, P. et al., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6), pp.626–635.
- Check Hayden, E., 2014. Technology: The $1,000 genome. *Nature*, 507(7492), pp.294–295.
- Cherbas, L. et al., 2011. The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Research*, 21(2), pp.301–314.
- Clark, A.G. et al., 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167), pp.203–218.

- Consortium, T.F., PMI, T.R. & DGT, C., 2014. A promoter-level mammalian expression atlas. *Nature*, 507(7493), pp.462–470.
- Core, L.J. et al., 2012. Defining the Status of RNA Polymerase at Promoters. *Cell Reports*, 2(4), pp.1025–1035.
- Core, L.J., Waterfall, J.J. & Lis, J.T., 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909), pp.1845–1848.
- Crocker, J., Ilsley, G.R. & Stern, D.L., 2016. Quantitatively predictable control of *Drosophila* transcriptional enhancers *in vivo* with engineered transcription factors. *Nature Genetics*, pp.–.
- Cusanovich, D.A. et al., 2014. The Functional Consequences of Variation in Transcription Factor Binding. *PLOS Genet*, 10(3), pp.e1004226 EP –.
- Dai, W., Zhang, G. & Makeyev, E.V., 2012. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Research*, 40(2), pp.787–800.
- Danecek, P. et al., The variant call format and VCFtools.
- Darren A Cusanovich, B.P.J.K.P.Y.G., 2014. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genetics*, 10(3).
- Das, A. et al., 2015. Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability. *Nature Communications*, 6, p.8555.
- Degner, J.F. et al., 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385), pp.390–394.
- Degner, J.F. et al., 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24), pp.3207–3212.
- Ding, Z. et al., 2014. Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association G. Gibson, ed. *PLOS Genet*, 10(11), p.e1004798.
- Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.
- Down, T.A. et al., 2007. Large-Scale Discovery of Promoter Motifs in *Drosophila* melanogaster. *PLOS Comput Biol*, 3(1), p.e7.
- Durrett, R. & Schmidt, D., 2008. Waiting for Two Mutations: With Applications to Regulatory Sequence Evolution and the Limits of Darwinian Evolution. *Genetics*, 180(3), pp.1501–1509.
- Elena, S.F. & Lenski, R.E., 1997. Test of synergistic interactions among deleterious mutations in bacteria : Article : Nature. *Nature*, 390(6658), pp.395–398.
- Erceg, J. et al., 2014. Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLOS Genet*, 10(1), p.e1004060.
- Flutre, T. et al., 2013. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLOS Genet*, 9(5), p.e1003486.
- Flynn, R.A. et al., 2016. 7SK-BAF axis controls pervasive transcription at enhancers. *Nature Structural & Molecular Biology*.

- Francesconi, M. & Lehner, B., 2013. The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482), pp.208–211.
- Franks, R.R. et al., 1990. Competitive titration in living sea urchin embryos of regulatory factors required for expression of the CyIIIa actin gene. *Development*, 110(1), pp.31–40.
- Gaffney, D.J. et al., 2012. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13(1), p.R7.
- Ghavi-Helm, Y. et al., 2014. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*.
- Gilad, Y., Rifkin, S.A. & Pritchard, J.K., 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, 24(8), pp.408–415.
- Gronau, I. et al., 2013. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. *Molecular Biology and Evolution*, 30(5), pp.1159–1171.
- Gulko, B. et al., 2014. Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. *bioRxiv*, p.006825.
- Gusella, J.F. et al., 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940), pp.234–238.
- Hammonds, A.S. et al., 2013. Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome Biology*, 14(12), p.1.
- Hare, E.E. et al., 2008. Sepsid even-skipped Enhancers Are Functionally Conserved in *Drosophila* Despite Lack of Sequence Conservation. *PLOS Genet*, 4(6), p.e1000106.
- Hertz, G.Z. & Stormo, G.D., Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. - PubMed - NCBI.
- Herzog, V.A. et al., 2014. A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nature Genetics*, 46(9), pp.973–981.
- Hilgers, V., Lemke, S.B. & Levine, M., 2012. ELAV mediates 3' UTR extension in the *Drosophila* nervous system. *Genes & Development*, 26(20), pp.2259–2264.
- Hill, W.G., Goddard, M.E. & Visscher, P.M., 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLOS Genet*, 4(2), pp.e1000008 EP –.
- Ho, J.W.K. et al., 2014. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515), pp.449–452.
- Hojoong Kwak, et al 2013. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science (New York, N.Y.)*, 339(6122), pp.950–953.
- Hoskins, R.A. et al., 2011. Genome-wide analysis of promoter architecture in *Drosophila* melanogaster. *Genome Research*, 21(2), pp.182–192.
- Huang, W. et al., 2012. Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. 109(39), pp.15553–15559.
- Jason Ernst, et al, 2010. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 20(4), p.526.

- Kharchenko et al., 2011. Comprehensive analysis of the chromatin landscape in *Drosophila* melanogaster. *Nature*, 471(7339), pp.480–485.
- Kvon, E.Z. et al., 2014. Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo* . *Nature*, pp.–.
- Kwak, H. et al., 2013. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science*, 339(6122), pp.950–953.
- Kwan, T. et al., 2008. Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics*, 40(2), pp.225–231.
- Lickwar, C.R. et al., 2012. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393), pp.251–255.
- Lin, S. et al., 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), pp.17224–17229.
- Lippert, C. et al., 2014. *LIMIX: genetic analysis of multiple traits*,
- Lynch, M., 2007. *The Origins of Genome Architecture*, Sinauer Associates Incorporated.
- Ma, X. et al., 2015. Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. *BMC Bioinformatics*, 16(1), pp.1–13.
- Mackay, T.F. & Moore, J.H., 2014. Why epistasis is important for tackling complex human disease genetics. *Genome medicine*, 6(6), p.124.
- Mackay, T.F.C., 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1), pp.22–33.
- Mathelier, A., Shi, W. & Wasserman, W.W., 2015. Identification of altered cis-regulatory elements in human disease. *Trends in genetics : TIG*, 31(2), pp.67–76.
- McDonald, J.H. & Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila* . *Nature*, 351(6328), pp.652–654.
- Mitra, S. & Narlikar, L., 2016. No Promoter Left Behind (NPLB): learn de novopromoter architectures from genome-wide transcription start sites. *Bioinformatics*, 32(5), pp.779–781.
- Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), pp.621–628.
- Mousavi, K. et al., 2013. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Molecular Cell*, 51(5), pp.606–617.
- Moyerbrailean, G.A. et al., 2016. Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLOS Genet*, 12(2), p.e1005875.
- Mustonen, V. & Lassig, M., 2007. Adaptations to fluctuating selection in *Drosophila* . *Proceedings of the National Academy of Sciences*, 104(7), pp.2277–2282.
- Myers, R.L. et al., 2007. Polymorphisms in the Regulatory Region of the Human Serotonin 5-HT2A Receptor Gene (HTR2A) Influence Gene Expression. *Biological Psychiatry*, 61(2), pp.167–173.

- Neve, J. et al., 2016. Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Research*, 26(1), pp.24–35.
- Nègre, N. et al., 2011. A cis-regulatory map of the *Drosophila* genome. *Nature*, 471(7339), pp.527–531.
- Ohler, U., 2006. Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Research*, 34(20), p.5943.
- Panne, D., 2008. The enhanceosome. *Current Opinion in Structural Biology*, 18(2), pp.236–242.
- Pelechano, V., Wei, W. & Steinmetz, L.M., 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447), pp.127–131.
- Philippe Batut, A.D.C.P.P.C.T.R.G., 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research*, 23(1), p.169.
- Pickrell, J.K. et al., 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289), pp.768–772.
- Pietrokovski, S., 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24(19), pp.3836–3845.
- Pique-Regi, R. et al., Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *genome.cshlp.org*.
- Proudfoot, N.J., 2011. Ending the message: poly(A) signals then and now. *Genes & Development*, 25(17), pp.1770–1782.
- Rach, E.A. et al., 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. *Genome Biology*, 10(7), pp.1–24.
- Rhee, D.Y. et al., 2014. Transcription Factor Networks in *Drosophila* melanogaster. *Cell Reports*, 8(6), pp.2031–2043.
- Rhee, H.S. & Pugh, B.F., 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6), pp.1408–1419.
- Rohs, R. et al., 2009. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268), pp.1248–1253.
- Sandmann, T. et al., 2007. A core transcriptional network for early mesoderm development in *Drosophila* melanogaster. *Genes & Development*, 21(4), pp.436–449.
- Sandmann, T. et al., 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Developmental cell*, 10(6), pp.797–807.
- Schadt, E.E. et al., 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929), pp.297–302.
- Schaub, M.A. et al., Linking disease associations with regulatory information in the human genome.
- Schaukowitch, K. et al., 2014. Enhancer RNA facilitates NELF release from immediate early genes. *Molecular Cell*, 56(1), pp.29–42.

- Schena, M. et al., 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235), pp.467–470.
- Semotok, J.L. & Lipshitz, H.D., 2007. Regulation and function of maternal mRNA destabilization during early *Drosophila* development. *Differentiation*, 75(6), pp.482–506.
- Shim, H. & Stephens, M., 2015. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *The Annals of Applied Statistics*, 9(2), pp.665–686.
- Shiraki, T. et al., 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26), pp.15776–15781.
- Siepel, A., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8), pp.1034–1050.
- Sigova, A.A. et al., 2015. Transcription factor trapping by RNA in gene regulatory elements. *Science*, 350(6263), pp.978–981.
- Simcha, D., Price, N.D. & Geman, D., 2012. The Limits of De Novo DNA Motif Discovery. *PLoS ONE*, 7(11), p.e47836.
- Smibert, P. et al., 2012. Global Patterns of Tissue-Specific Alternative Polyadenylation in *Drosophila* . *Cell Reports*, 1(3), pp.277–289.
- Spivakov, M. et al., 2012. Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biology*, 13(9), p.R49.
- Sun, Y. et al., Zelda overcomes the high intrinsic nucleosome barrier at enhancers during *Drosophila* zygotic genome activation. *genome.cshlp.org*.
- Tajima, F., 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3), p.585.
- Takahashi, H. et al., 2012. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature protocols*, 7(3), pp.542–561.
- Trapnell, C., 2015. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10), pp.1491–1498.
- Tuan, D., Kong, S. & Hu, K., 1992. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proceedings of the National Academy of Sciences of the United States of America*, 89(23), pp.11219–11223.
- Veyrieras, J.-B. et al., 2008. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLOS Genet*, 4(10), p.e1000214.
- Ville Mustonen, M.L., 2007. Adaptations to fluctuating selection in *Drosophila* . *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), p.2277.
- Wickham, H. & Francois, R., 2014. dplyr: A grammar of data manipulation. *URL http://CRAN. R-project. org/package= dplyr. R package version 0.2*.
- William G Hill, M.E.G.P.M.V., 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics*, 4(2).
- Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3), pp.206–216.

- Wu, H. et al., 2014. Tissue-Specific RNA Expression Marks Distant-Acting Developmental Enhancers. *PLOS Genet*, 10(9), p.e1004610.
- Yao, P. et al., Coexpression networks identify brain region-specific enhancer RNAs in the human brain. 18(8), pp.1168–1174.
- Yoon, O.K. et al., 2012. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLOS Genet*, 8(8), p.e1002882.
- Young, R.S. et al., 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Research*, 25(10), pp.1546–1557.

Appendix

Table 3.2 - Discovered Promoter Motifs

simcluster	centrimo_bin_location	fitz_Matches	Tiffin_Matches	is_novel	shape_shift	prevalence
1	-0.5	GGYCACAC	TIFDMEM0000116	FALSE	-0.403172645	0.267521248
1	-1.5	GGYCACAC	TIFDMEM0000116	FALSE	-0.494971589	0.096837119
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.490044459	0.089243416
8	30.5	NA	NA	TRUE	-0.072126086	0.019088756
8	185.5	NA	NA	TRUE	-0.211481948	0.004110353
11	-26.5	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.681479878	0.019994427
2	NA	NA	TIFDMEM0000002	FALSE	-0.103791746	0.002299011
5	1.5	CAGCTSWW	TIFDMEM0000079	FALSE	0	0.002299011
5	1.5	CAGCTSWW	TIFDMEM0000079	FALSE	-0.052986104	0.022014769
2	NA	NA	TIFDMEM0000002	FALSE	-0.256918046	0.023826111
10	0	TCAGTY	TIFDMEM0000077	FALSE	1.194998657	0.023826111
10	0.5	TCAGTY	TIFDMEM0000077	FALSE	1.267026685	0.026752125
3	4.5	NA	TIFDMEM0000076;TIFDMEM0000101	FALSE	0.052777279	0.793576703
6	16.5	NA	TIFDMEM0000065	FALSE	-0.070165777	0.744113139
8	26.5	NA	NA	TRUE	-0.27368247	0.739027449
8	NA	NA	NA	TRUE	0.076968868	0.150341368
11	-25.5	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.641034656	0.587362408
12	22.5	NA	NA	FALSE	0.429228541	0.458199805
4	242.5	NA	NA	TRUE	-0.193485947	0.248432493
4	241.5	NA	NA	TRUE	-0.15590459	0.349797966
2	NA	NA	TIFDMEM0000002	FALSE	-0.147002212	0.218754354
4	244.5	NA	NA	TRUE	-0.162218611	0.075310018

6	34.5	NA	TIFDMEM0000065	FALSE	-0.09279238	0.089870419
13	17.5	CGMYGYCR	NA	FALSE	0.100833945	0.086665738
5	1.5	CAGCTSWW	TIFDMEM0000079	FALSE	-0.030449566	0.124076912
7	6.5	CARCCCT	TIFDMEM0000042;TIFDMEM0000109	FALSE	-0.255801388	0.338093911
9	-5.5	NA	NA	FALSE	-0.262058114	0.261320886
6	2.5	NA	TIFDMEM0000065	FALSE	-0.191680446	0.265361572
12	20.5	NA	NA	FALSE	0.399573155	0.110561516
14	28.5	NA	NA	TRUE	0.330649986	0.048209558
29	0	NA	TIFDMEM0000080	FALSE	-0.249123695	0.581301379
51	0	NA	TIFDMEM0000072;TIFDMEM0000084;TIFDMEM0000094	FALSE	0.577790769	0.55420092
29	0	NA	TIFDMEM0000080	FALSE	-0.036034603	0.317124147
6	30	NA	TIFDMEM0000065	FALSE	-0.36678572	0.041730528
29	27.5	NA	TIFDMEM0000080	FALSE	0.299364141	0.103734151
6	NA	NA	TIFDMEM0000065	FALSE	-0.14753647	0.041521527
18	NA	NA	NA	TRUE	-0.276669039	0.011843389
29	NA	NA	TIFDMEM0000080	FALSE	0	0.045562213
2	NA	NA	TIFDMEM0000002	FALSE	-0.315532495	0.03497283
20	NA	TGGTATTT	TIFDMEM0000073;TIFDMEM0000091;TIFDMEM0000111	FALSE	-0.379313279	0.316706145
48	NA	NA	TIFDMEM0000055	FALSE	-0.465008938	0.16650411
20	-25	TGGTATTT	TIFDMEM0000073;TIFDMEM0000091;TIFDMEM0000111	FALSE	-0.638249879	0.069527658
6	6	NA	TIFDMEM0000065	FALSE	-0.589673112	0.058311272
6	6	NA	TIFDMEM0000065	FALSE	-0.503433478	0.380103107
29	5.5	NA	TIFDMEM0000080	FALSE	0.152995342	0.535460499
4	NA	NA	NA	TRUE	0	0.121638568
33	1	NA	TIFDMEM0000038;TIFDMEM0000062	FALSE	-0.904799322	0.647485022
27	NA	NA	NA	TRUE	0.129532248	0.196530584
33	NA	NA	TIFDMEM0000038;TIFDMEM0000062	FALSE	0	0.043820538

33	NA	NA	TIFDMEM0000038;TIFDMEM0000062	FALSE	0	0.050856904
1	-3	GGYCACAC	TIFDMEM0000116	FALSE	-0.819830803	0.01504807
4	NA	NA	NA	TRUE	0	0.011843389
31	-0.5	NA	TIFDMEM0000058	FALSE	0.837099833	0.160512749
10	-0.5	TCAGTY	TIFDMEM0000077	FALSE	1.087005618	0.155078724
34	NA	NA	NA	TRUE	0	0.20426362
5	-6	CAGCTSWW	TIFDMEM0000079	FALSE	-0.551528552	0.158353072
21	NA	NA	NA	TRUE	-0.377763416	0.184269193
35	-80	GAGAGCG	TIFDMEM0000009	FALSE	0.705376254	0.006966699
21	NA	NA	NA	TRUE	-0.323808483	0.160861084
6	26	NA	TIFDMEM0000065	FALSE	-0.041846768	0.166086108
24	-0.5	NA	NA	TRUE	0	0.034763829
7	2.5	CARCCCT	TIFDMEM0000042;TIFDMEM0000109	FALSE	-0.470099583	0.08847708
39	-12	NA	NA	TRUE	-0.410265653	0.132576285
29	2.5	NA	TIFDMEM0000080	FALSE	0.233442394	0.113278529
18	29	NA	NA	TRUE	-0.412297197	0.113278529
52	NA	NA	NA	TRUE	0.272378429	0.389090149
18	NA	NA	NA	TRUE	-0.28604228	0.427824997
36	NA	GAAAGCT	TIFDMEM0000018	FALSE	0.098651392	0.449003762
50	NA	NA	TIFDMEM0000021;TIFDMEM0000034;TIFDMEM0000068;TIFDMEM0000095	FALSE	-0.196204899	0.159049742
46	18	NA	NA	TRUE	0.143784432	0.374460081
27	NA	NA	NA	TRUE	0.175249668	0.055524592
20	-25	TGGTATTT	TIFDMEM0000073;TIFDMEM0000091;TIFDMEM0000111	FALSE	-0.551942084	0.058102271
54	NA	NA	TIFDMEM0000017;TIFDMEM0000024	FALSE	-0.316868942	0.217012679
5	3	CAGCTSWW	TIFDMEM0000079	FALSE	-0.866718768	0.117597882
48	NA	NA	TIFDMEM0000055	FALSE	-0.196504092	0.314685802
23	-3	NA	NA	TRUE	0.280023776	0.362477358

6	27.5	NA	TIFDMEM0000065	FALSE	-0.079415975	0.149575031
31	0.5	NA	TIFDMEM0000058	FALSE	0.851960526	0.134666295
15	1	NA	TIFDMEM0000015	FALSE	1.05517167	0.167967117
9	-4	NA	NA	FALSE	-0.644164342	0.187055873
4	NA	NA	NA	TRUE	0	0.160094747
47	30.5	CGGACGT	NA	FALSE	-0.237614718	0.282917654
45	1.5	CAYCNCTA	NA	FALSE	-0.286769487	0.282917654
34	NA	NA	NA	TRUE	0	0.720356695
7	2	CARCCCT	TIFDMEM0000042;TIFDMEM0000109	FALSE	-0.756093147	0.450327435
27	NA	NA	NA	TRUE	0.219649825	0.887975477
15	171.5	NA	TIFDMEM0000015	FALSE	-0.22991786	0.278040964
28	19	NA	TIFDMEM0000057	FALSE	0.081875377	0.96063815
7	2	CARCCCT	TIFDMEM0000042;TIFDMEM0000109	FALSE	-0.78855298	0.740978125
15	1	NA	TIFDMEM0000015	FALSE	0.532891883	0.068970322
15	2	NA	TIFDMEM0000015	FALSE	0.101530021	0.596837119
15	1	NA	TIFDMEM0000015	FALSE	0.100965075	0.386721471
9	-4	NA	NA	FALSE	-0.659878288	0.02919047
9	-4	NA	NA	FALSE	-0.654736493	0.012470392
46	-37	NA	NA	TRUE	-0.5792243	0.08429706
9	-4	NA	NA	FALSE	-0.347207276	0.083530723
34	2	NA	NA	TRUE	-0.509179956	0.056708931
41	NA	NA	NA	TRUE	0	0.078584367
53	NA	NA	TIFDMEM0000117	FALSE	0.031857346	0.034554828
41	NA	NA	NA	TRUE	0	0.046398217
39	14	NA	NA	TRUE	-0.395921328	0.038038178
40	1	TCATTCG	TIFDMEM0000102	FALSE	-0.351143465	0.028006131
58	NA	NA	NA	TRUE	0.340779414	0.042357531

26	19.5	NA	TIFDMEM0000010	FALSE	0.085849743	0.50996238
57	26	KCGGTTSK	NA	FALSE	0.409674683	0.216176675
40	24	TCATTCG	TIFDMEM0000102	FALSE	0.525300486	0.213598997
28	-2	NA	TIFDMEM0000057	FALSE	0	0.266824579
34	14.5	NA	NA	TRUE	0.296969224	0.029956806
59	24	NA	NA	TRUE	0.516068767	0.036505504
37	-42.5	NA	TIFDMEM0000059;TIFDMEM0000096	FALSE	-0.704513029	0.514003065
17	-44	NA	TIFDMEM0000074	FALSE	-0.644120273	0.096488784
17	-44	NA	TIFDMEM0000074	FALSE	-0.568534753	0.239375784
17	-44	NA	TIFDMEM0000074	FALSE	-0.566888703	0.028006131
47	29	CGGACGT	NA	FALSE	0.849435411	0.092308764
13	20.5	CGMYGYCR	NA	FALSE	0.212639337	0.086247736
26	19	NA	TIFDMEM0000010	FALSE	0.365811527	0.458826808
49	NA	NA	NA	TRUE	0.124128548	0.256513864
26	19	NA	TIFDMEM0000010	FALSE	0.177308428	0.293089034
56	18.5	NA	NA	TRUE	0.229157459	0.014003065
21	NA	NA	NA	TRUE	0	0.087919744
24	NA	NA	NA	TRUE	0.024057056	0.243346802
24	21	NA	NA	TRUE	0.343633852	0.041103525
35	-77	GAGAGCG	TIFDMEM0000009	FALSE	0.747113674	0.194579908
24	26	NA	NA	TRUE	0.14350769	0.047443221
30	-2	NA	TIFDMEM0000040	FALSE	-0.30316918	0.187264874
30	-2	NA	TIFDMEM0000040	FALSE	-0.565308672	0.161697088
21	21.5	NA	NA	TRUE	-0.373017394	0.137452975
11	-27	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	1.400951499	0.103734151
24	24.5	NA	NA	TRUE	0.187982621	0.003344016
15	-2	NA	TIFDMEM0000015	FALSE	-0.234350053	0.106311829

18	NA	NA	NA	TRUE	-0.340763884	0.109168176
10	0	TCAGTY	TIFDMEM0000077	FALSE	1.307332167	0.170266128
22	NA	NA	NA	TRUE	0.066784166	0.012749059
22	223.5	NA	NA	TRUE	-0.035076898	0.011634388
21	NA	NA	NA	TRUE	-0.321378665	0.035112164
32	0	NA	NA	TRUE	0	0.027379128
14	NA	NA	NA	TRUE	0	0.010032047
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.429848155	0.003344016
43	-11	NA	NA	TRUE	0.045038038	0.011146719
38	-20	NA	TIFDMEM0000063	FALSE	0.146360421	0.001463007
25	NA	NA	TIFDMEM0000012;TIFDMEM0000046;TIFDMEM0000069;TIFDMEM0000090;TIFDMEM0000093	FALSE	0	0.020273095
19	-5	TGGYAACR	TIFDMEM0000026	FALSE	-0.504748178	0.061585621
22	NA	NA	NA	TRUE	0.041142157	0.051205239
22	NA	NA	NA	TRUE	0	0.039083182
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.845304891	0.059774279
1	0	GGYCACAC	TIFDMEM0000116	FALSE	-0.217209832	0.031001811
1	0	GGYCACAC	TIFDMEM0000116	FALSE	-0.243706651	0.054967257
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.847160474	0.075379685
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.834981648	0.046049882
1	-1	GGYCACAC	TIFDMEM0000116	FALSE	-0.83019025	0.072453671
55	NA	NA	NA	TRUE	0	0.329733872
27	NA	NA	NA	TRUE	-0.03996171	0.062212624
19	-2	TGGYAACR	TIFDMEM0000026	FALSE	-0.537436227	0.016232409
18	-3.5	NA	NA	TRUE	-0.423891453	0.591124425
19	-3.5	TGGYAACR	TIFDMEM0000026	FALSE	-0.218640818	0.008429706
16	-4	ATCGATA	TIFDMEM0000005;TIFDMEM0000098;TIFDMEM0000118	FALSE	-0.827005697	0.008151038
16	-46	ATCGATA	TIFDMEM0000005;TIFDMEM0000098;TIFDMEM0000118	FALSE	-0.767343472	0.121220566

23	-2.5	NA	NA	TRUE	0.226506149	0.086665738
56	NA	NA	NA	TRUE	0	0.442942734
25	-3.5	NA	TIFDMEM0000012;TIFDMEM0000046;TIFDMEM0000069;TIFDMEM0000090;TIFDMEM0000093	FALSE	-0.001480515	0.730737077
58	NA	NA	NA	TRUE	0.050603084	0.215480006
4	NA	NA	NA	TRUE	0	0.851051972
10	0	TCAGTY	TIFDMEM0000077	FALSE	1.256536036	0.022990107
43	NA	NA	NA	TRUE	-0.166335991	0.031141145
2	NA	NA	TIFDMEM0000002	FALSE	-0.317761871	0.026334123
43	NA	NA	NA	TRUE	0.369163765	0.015814407
54	NA	NA	TIFDMEM0000017;TIFDMEM0000024	FALSE	-0.366364633	0.012331058
32	-1	NA	NA	TRUE	-0.37535218	0.051623241
44	NA	NA	NA	TRUE	-0.130231003	0.620454229
11	-28.5	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.880545664	0.380451442
39	NA	NA	NA	TRUE	0.030134576	0.213389996
10	-0.5	TCAGTY	TIFDMEM0000077	FALSE	1.125007436	0.022432771
11	-28	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.636455375	0.102480145
11	-27.5	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.743770731	0.070154661
11	-27.5	STATAAA	TIFDMEM0000083;TIFDMEM0000107	FALSE	0.73790439	0.420718963
16	-41.5	ATCGATA	TIFDMEM0000005;TIFDMEM0000098;TIFDMEM0000118	FALSE	-0.779697456	0.492893967
17	17.5	NA	TIFDMEM0000074	FALSE	-0.010889171	0.06785565
25	NA	NA	TIFDMEM0000012;TIFDMEM0000046;TIFDMEM0000069;TIFDMEM0000090;TIFDMEM0000093	FALSE	0	0.36568204
23	NA	NA	NA	TRUE	0.07947721	0.087014073
18	NA	NA	NA	TRUE	-0.304045301	0.211996656
18	NA	NA	NA	TRUE	-0.286337217	0.065347638
17	-40	NA	TIFDMEM0000074	FALSE	-0.393301721	0.677232827
1	-2	GGYCACAC	TIFDMEM0000116	FALSE	-0.748611146	0.406855232

simcluster	MotifName	tssset	E_value	iupac	ohlermat	InformationContent	tomtomMatch	locationset	centrimo_E_value	centrimo_bin_width
1	2	all_tss	2.9E-61	YGGTCACACTG	Motif_1.ohler	13.89666867	FBgn0027339_2	TSS	2.50E-236	64
1	2	broad_tss	1.80E-171	YGGTCACACTG	Motif_1.ohler	14.17395822	FBgn0001994	TSS	7.50E-235	4
1	2	broad_v_narrow	3.20E-166	YGGTCACACTG	Motif_1.ohler	14.15150872	FBgn0001994	TSS	2.00E-282	3
8	2	internal_tss	1.2E-57	AGCARCARCARCAGC	NA	14.53274234	FBgn0027339_2	Downstream	0.00014	94
8	2	narrow_tss	4.8E-22	CRGCARCAGCAGCAR	NA	12.20448681	NA	Downstream	3.7E-09	126
11	2	narrow_v_broad	9.4E-28	TATAWAAR	TATA.ohler	12.07081904	NA	Upstream	2.3E-64	62
2	3	all_tss	3.6E-23	AAAYAAMAAMAMAA	NA	12.39679803	NA	NA	NA	NA
5	3	broad_tss	2.2E-29	CVRWCAGCTGTT	NA	12.19910141	FBgn0038787_2	TSS	2.20E-119	68
5	3	broad_v_narrow	2.2E-30	SMRAWCAGCTGTT	NA	12.62476204	NA	TSS	1.2E-98	68
2	3	internal_tss	4.5E-27	WAMWWMAWAMAHAAA	NA	12.52438686	NA	NA	NA	NA
10	3	narrow_tss	3.3E-34	TCAGTYK	INR.ohler	10.68649012	NA	TSS	0	3
10	3	narrow_v_broad	1.8E-34	TCAGTYK	INR.ohler	10.53078085	NA	TSS	0	4
3	4	all_tss	4.3E-18	CASCRRCARC	NA	10.42888365	FBgn0003499_2	TSS	8.6E-35	102
6	4	broad_tss	2.3E-22	SMRMARSARMAVVAR	NA	11.21696208	FBgn0013469	Downstream	0.0000031	136
8	4	broad_v_narrow	2.2E-17	SAGCRRSARMARSAR	NA	11.79424522	FBgn0027339_2	Downstream	0.042	2
8	4	internal_tss	4.5E-19	SAGSWGGAGRHGSWG	NA	11.08972744	RNCMPT00283	NA	NA	NA
11	4	narrow_tss	9.2E-27	STATAWAAR	TATA.ohler	12.21054362	FBgn0005694_2	Upstream	1.3E-62	62
12	4	narrow_v_broad	0.0048	CGHKBCGYTCG	MTE.ohler	10.76248046	NA	Downstream	9.6E-97	92
4	5	all_tss	5.8E-16	GYGWGYGTGTGTGTG	NA	11.79395133	FBgn0002922_2	Downstream	0.4	68
4	5	broad_tss	5.3E-12	GTGTSTGYGTSKGTG	NA	11.4371153	NA	Downstream	1.4	72
2	5	broad_v_narrow	1.2E-20	AAAAMDAAMMAMA	NA	13.10532368	NA	NA	NA	NA
4	5	internal_tss	0.25	GTGTGYGTGTG	NA	10.02659335	Aef1	Downstream	6.8	60
6	5	narrow_tss	0.000001	GMMAMAAMAAMAA	NA	12.08390223	NA	Downstream	0.0014	162
13	5	narrow_v_broad	2.9	CRGCGRCRGC	NA	10.11207302	NA	Downstream	3.8E-45	104
5	6	all_tss	0.00031	RTCAGCTGTT	NA	10.91463905	NA	TSS	2.50E-116	64
7	6	broad_tss	9.6E-11	GYTKTRCARCACTG	NA	12.16851601	NA	TSS	1.7E-45	124

9	6	broad_v_narrow	3E-17	YCATCHCTA	Motif_7.ohler	11.99065754	NA	TSS	3.4E-56	64
6	6	internal_tss	1.3	AACARCAACAACAR	NA	11.11953471	NA	TSS	0.34	98
12	6	narrow_tss	0.00085	CKCKSMGCTCGS	MTE.ohler	10.48876695	NA	Downstream	8.4E-87	92
14	6	narrow_v_broad	5900	MGMHGCGRACGMRYK	NA	9.330553804	NA	Downstream	5E-96	82
29	AAACDAAA	broad_tss	0.000048	AAACDAAA	NA	13.65575083	NA	TSS	2.7E-16	125
51	AAACHRAA	internal_tss	9.1E-35	TTYDGTTT	NA	13.43054805	RNCMPT00112	TSS	3.9E-49	3
29	AAAYGAAA	all_tss	0.027	TTTCRTTT	NA	14.33193664	NA	TSS	0.000000022	3
6	AACAACAV	narrow_tss	0.0000028	AACAACAV	NA	13.41558271	FBgn0005694_2	Downstream	4.3E-09	69
29	AACSGAA	all_tss	1.3E-10	AACSGAA	NA	12.63676693	NA	Downstream	1.2E-10	204
6	AACWACAA	internal_tss	1.10E-125	AACWACAA	NA	15.15544934	FBgn0038787_2	NA	NA	NA
18	AAHAAW	broad_v_narrow	7.90E-200	AAHAAW	NA	9.419385433	NA	NA	NA	NA
29	AANGGAAA	internal_tss	2.4E-16	AANGGAAA	NA	13.79623055	NA	NA	NA	NA
2	AATAABAA	internal_tss	2.9E-21	AATAABAA	NA	14.17133653	NA	NA	NA	NA
20	AATACCA	all_tss	0.000017	TGGTATT	Motif_6.ohler	12.76888547	NA	NA	NA	NA
48	AATATAYC	broad_v_narrow	0.033	GRTATATT	NA	11.09620776	NA	NA	NA	NA
20	AAWATACC	all_tss	6.1E-46	GGTATWTT	Motif_6.ohler	14.34639723	NA	Upstream	0.000027	185
6	AAYAACAA	broad_tss	2.9E-33	AAYAACAA	NA	14.44366212	FBgn0038787_2	TSS	0.000078	69
6	AAYAAYAA	all_tss	5.6E-54	TTRTTRTT	NA	13.87148727	FBgn0005694_2	TSS	2.6	5
29	AAYSGAA	narrow_tss	9.6E-13	AAYSGAA	NA	11.67438486	NA	TSS	1.9E-11	142
4	ACACMCAC	internal_tss	4.8E-66	GTGKGTGT	NA	14.98849606	RNCMPT00011	NA	NA	NA
33	ACAGATGW	broad_v_narrow	0.000093	ACAGATGW	NA	11.69754878	NA	TSS	2.4E-35	13
27	ACASATAC	broad_tss	0.0073	ACASATAC	NA	12.62802817	NA	NA	NA	NA
33	ACATATGT	broad_tss	0.000048	ACATATGT	NA	12.31498134	NA	NA	NA	NA
33	ACATATRT	all_tss	0.000049	ACATATRT	NA	13.03616985	NA	NA	NA	NA
1	ACGGTCAC	broad_tss	6.5E-13	ACGGTCAC	Motif_1.ohler	13.5170049	NA	TSS	9.80E-140	59
4	ACRCACAC	broad_tss	3E-21	GTGTGYGT	NA	14.27612819	RNCMPT00011	NA	NA	NA
31	ACTAAAH	all_tss	0.000000015	DTTTAGT	NA	12.14277673	NA	TSS	5.60E-180	4

10	ACTGR	all_tss	2.1E-72	YCAGT	INR.ohler	9.005680056	NA	TSS	0	2
34	ACTYACC	internal_tss	1.1E-55	GGTRAGT	NA	12.80745951	NA	NA	NA	NA
5	ADCAGCWG	broad_tss	3.40E-170	ADCAGCWG	NA	13.31696805	FBgn0002922_2	TSS	4.8E-72	91
21	AGAAGAAG	all_tss	2.1E-23	AGAAGAAG	NA	14.93936531	RNCMPT00078	NA	NA	NA
35	AGAGRGAG	all_tss	0.0093	AGAGRGAG	NA	13.56295604	NA	Upstream	0.049	245
21	AGARGAAG	broad_v_narrow	1.2E-17	AGARGAAG	NA	14.3037607	RNCMPT00078	NA	NA	NA
6	AGCAACAR	all_tss	0.023	AGCAACAR	NA	14.10570357	NA	Downstream	0.000012	153
24	AGCAGWG	broad_tss	0.00037	AGCAGWG	NA	12.13725113	NA	TSS	3E-32	220
7	AGDGNTG	broad_v_narrow	6.5E-99	CANCHCT	NA	10.41797037	NA	TSS	7.1E-84	70
39	AGRTGG	broad_v_narrow	0.0061	CCAYCT	NA	10.61666693	NA	Upstream	5.6	61
29	AGTSGAA	all_tss	0.032	AGTSGAA	NA	12.28787282	NA	TSS	2.6E-18	14
18	AHAACA	broad_v_narrow	1.8E-09	AHAACA	NA	10.40397461	NA	Downstream	0.000023	73
52	AKCSGGA	internal_tss	4.4E-31	AKCSGGA	NA	11.89309159	NA	NA	NA	NA
18	AMATW	broad_v_narrow	3.6E-56	AMATW	NA	8.021374147	NA	NA	NA	NA
36	ASTTTTCC	all_tss	0.023	ASTTTTCC	NA	13.16969041	NA	NA	NA	NA
50	ATAYAHAT	internal_tss	7.8E-98	ATAYAHAT	NA	13.3129457	NA	NA	NA	NA
46	ATCGCDGC	narrow_tss	0.0015	ATCGCDGC	NA	12.68299773	NA	Downstream	1.1E-50	157
27	ATCYGHA	internal_tss	2.1E-21	ATCYGHA	NA	11.3848137	NA	NA	NA	NA
20	AWATACCR	broad_tss	2.8E-47	YGGTATWT	Motif_6.ohler	13.57545621	NA	Upstream	6.6	197
54	AWATATW	internal_tss	4E-23	AWATATW	NA	12.04344957	NA	NA	NA	NA
5	AWCAGCTG	broad_v_narrow	4.9E-81	AWCAGCTG	NA	14.64252272	FBgn0002922_2	TSS	3.1E-73	65
48	AWVTAC	internal_tss	3.1E-88	GTABWT	NA	9.406921383	NA	NA	NA	NA
23	BYTGGCCA	all_tss	1.2E-19	BYTGGCCA	NA	12.75797451	NA	TSS	0.0000017	51
6	CAACAAC	internal_tss	7.9E-19	CAACAAC	NA	13.72229174	Aef1	Downstream	0.000045	68
31	CAACTAA	narrow_v_broad	0.046	TTAGTTG	NA	12.55784747	NA	TSS	0	4
15	CAACYG	narrow_v_broad	2.2E-13	CRGTTG	NA	10.75545222	NA	TSS	0	5
9	CABCWCTA	all_tss	1.7E-71	CABCWCTA	Motif_7.ohler	13.13467125	NA	TSS	1.10E-136	65

4	CACACACR	all_tss	2.7E-33	YGTGTGTG	NA	14.4512021	RNCMPT00283	NA	NA	NA
47	CACGTGY	broad_v_narrow	0.018	CACGTGY	NA	11.37626746	HLHm5	Downstream	5.1E-26	66
45	CACSCTR	broad_v_narrow	0.0000032	CACSCTR	NA	11.18175264	NA	TSS	3.9E-32	84
34	CACYCACA	all_tss	0.02	TGTGRGTG	NA	13.29652763	NA	NA	NA	NA
7	CADCMCTG	broad_tss	7.2E-99	CADCMCTG	NA	13.20034084	NA	TSS	2.4E-91	121
27	CAGATAC	all_tss	1.3E-11	CAGATAC	NA	12.99627693	NA	NA	NA	NA
15	CAGCDVC	internal_tss	2.70E-284	CAGCDVC	NA	10.87701432	NA	Downstream	1.7E-12	100
28	CAGCSA	broad_tss	0.0026	CAGCSA	NA	10.74809813	NA	Downstream	6.6E-20	95
7	CAGKGYTG	all_tss	1.4E-31	CARCMCTG	NA	13.36147137	NA	TSS	5.1E-62	119
15	CAGYBG	broad_tss	1.8E-19	CAGYBG	NA	9.355302462	NA	TSS	1.20E-218	5
15	CAGYHG	all_tss	4.50E-251	CAGYHG	NA	9.404256972	NA	TSS	3.10E-129	51
15	CAGYNG	narrow_tss	6.9E-88	CAGYNG	NA	8.962903486	NA	TSS	2.6E-70	51
9	CAHCHCTA	broad_tss	2.8E-96	CAHCHCTA	Motif_7.ohler	12.68869963	NA	TSS	1.70E-140	65
9	CATCCCTR	all_tss	7.2E-23	CATCCCTR	Motif_7.ohler	13.65027279	NA	TSS	1.3E-37	115
46	CATCGATK	broad_v_narrow	0.0051	CATCGATK	NA	13.05241897	NA	Upstream	4E-23	157
9	CAWCGCTA	broad_tss	0.00000016	CAWCGCTA	Motif_7.ohler	12.95831432	NA	TSS	3.6E-66	127
34	CAYACH	broad_v_narrow	1E-17	CAYACH	NA	9.647216738	RNCMPT00285	TSS	2.8E-86	5
41	CAYCAYC	internal_tss	2.5E-31	CAYCAYC	NA	11.96426259	NA	NA	NA	NA
53	CCABTKCC	internal_tss	2.1E-25	CCABTKCC	NA	13.23243668	NA	NA	NA	NA
41	CCACCAS	broad_tss	0.000016	CCACCAS	NA	12.00511033	NA	NA	NA	NA
39	CCASCTCY	broad_tss	3.1E-11	CCASCTCY	NA	12.60108264	NA	Downstream	0.00000017	63
40	CCATTCRC	broad_tss	0.0000091	CCATTCRC	NA	13.14085175	NA	TSS	2.3E-26	53
58	CCCCMAA	narrow_tss	0.00055	CCCCMAA	NA	11.79169959	NA	NA	NA	NA
26	CCGGMGA	narrow_tss	0.015	CCGGMGA	NA	11.97688073	NA	Downstream	1E-17	50
57	CCGMWC	narrow_tss	9.1E-31	GWKCGG	DPE.ohler	9.887181488	NA	Downstream	1.8E-42	101
40	CCGTTMGC	narrow_tss	0.0048	GCKAACGG	NA	12.45513726	NA	Downstream	1.4E-72	61
28	CDGCCA	all_tss	5.4E-11	CDGCCA	NA	10.35767577	NA	TSS	1.1E-18	51

34	CGAGTGB	all_tss	0.00019	CGAGTGB	NA	11.94512263	NA	Downstream	0.000027	52
59	CGANCB	narrow_v_broad	1.40E-170	VGNTCG	NA	8.482589029	NA	Downstream	2.20E-116	99
37	CGATAAC	all_tss	0.048	CGATAAC	NA	12.31742459	NA	Upstream	4.4E-88	178
17	CGATAG	all_tss	0.00000024	CGATAG	NA	11.37142305	NA	Upstream	7.20E-159	163
17	CGATAR	broad_tss	1.9E-77	CGATAR	NA	10.84710866	NA	Upstream	1.20E-162	177
17	CGATAR	broad_v_narrow	1.4E-52	CGATAR	NA	10.81927814	NA	Upstream	1.20E-162	177
47	CGGACGYG	narrow_tss	0.000091	CGGACGYG	NA	12.21648091	NA	Downstream	8.6E-61	85
13	CGGCRGM	narrow_tss	1.1E-13	CGGCRGM	NA	11.56881545	Mad	Downstream	2.4E-14	190
26	CGRCGA	all_tss	3.9E-12	TCGYCG	NA	10.8502835	NA	Downstream	1.9E-40	51
49	CRASGA	internal_tss	1.20E-245	CRASGA	NA	9.988275645	NA	NA	NA	NA
26	CRGCGA	internal_tss	6.6E-51	TCGCYG	NA	10.95534347	NA	Downstream	8.6E-32	161
56	CSAAG	narrow_v_broad	0.000016	CSAAG	NA	8.87029038	NA	Downstream	0.000000078	116
21	CTACTWC	internal_tss	1E-12	CTACTWC	NA	12.59979525	NA	NA	NA	NA
24	CTBCTSC	internal_tss	0	GSAGVAG	NA	11.52384893	NA	NA	NA	NA
24	CTCKGCTC	all_tss	0.004	CTCKGCTC	NA	12.97957526	NA	Downstream	3.6E-28	219
35	CTCTCTCB	broad_tss	0.00014	CTCTCTCB	NA	12.6435798	NA	Upstream	4.2E-10	251
24	CTGCTC	narrow_tss	0.044	GAGCAG	NA	11.4546028	NA	Downstream	0.00000026	101
30	CTGDTAAC	all_tss	0.000000009	CTGDTAAC	NA	12.83928101	NA	TSS	2.7E-30	61
30	CTGGYMAC	broad_tss	6.4E-37	CTGGYMAC	NA	13.25606462	NA	TSS	1.7E-80	51
21	CTTCTTS	internal_tss	9.4E-81	SAAGAAG	NA	12.89033015	RNCMPT00078	Downstream	0.000000022	50
11	CTTWTATA	narrow_v_broad	0.0000091	TATAWAAG	TATA.ohler	12.58739712	NA	Upstream	2.8	63
24	CWGCTCS	all_tss	1.5E-14	SGAGCWG	NA	11.70515006	NA	Downstream	0.0000002	100
15	CWGCWG	internal_tss	5.8E-49	CWGCWG	NA	10.02610782	NA	TSS	4.8E-26	51
18	DAAATA	broad_tss	8.1E-26	DAAATA	NA	10.37578412	NA	NA	NA	NA
10	DACTGA	narrow_v_broad	8.00E-303	TCAGTH	INR.ohler	10.5770494	NA	TSS	0	3
22	DCCRCC	all_tss	2.8E-23	DCCRCC	NA	9.360180198	NA	NA	NA	NA
22	DCCRCCR	internal_tss	0	YGGYGGH	NA	10.43256868	NA	Downstream	6.5	98

21	GAAGARGA	broad_tss	1.3E-26	GAAGARGA	NA	14.09501962	NA	NA	NA	NA
32	GAATGSCA	all_tss	0.000017	TGSCATTC	NA	13.3751798	NA	TSS	1.3E-23	3
14	GADGMGGA	internal_tss	3.1E-47	GADGMGGA	NA	13.23252153	NA	NA	NA	NA
1	GAGTGACC	broad_v_narrow	0.000016	GGTCACTC	Motif_1.ohler	13.78598255	NA	TSS	5.50E-260	3
43	GCCAVM	narrow_tss	2E-34	GCCAVM	NA	9.37963761	NA	Upstream	2.3E-25	109
38	GCYGSC	broad_tss	2.6E-46	GSCRGC	NA	9.967107084	NA	Upstream	6.1E-22	127
25	GGAAAW	broad_tss	0.000000079	WTTTCC	NA	10.9098239	dl	NA	NA	NA
19	GGCAACRC	broad_v_narrow	1.9E-10	GGCAACRC	NA	13.95608253	NA	TSS	4.4E-69	127
22	GGCGGHR	internal_tss	5E-29	GGCGGHR	NA	11.31973099	NA	NA	NA	NA
22	GGHGGA	broad_tss	0.00046	GGHGGA	NA	10.13739527	NA	NA	NA	NA
1	GGTCACAC	narrow_tss	0.00028	GGTCACAC	Motif_1.ohler	12.31498134	NA	TSS	0.00E+00	3
1	GGTCATAC	all_tss	0.00000019	GGTCATAC	Motif_1.ohler	13.01114329	NA	TSS	1.80E-230	3
1	GGTCATAC	broad_tss	1.9E-11	GGTCATAC	Motif_1.ohler	13.07483934	NA	TSS	1.80E-230	3
1	GGTMACAC	broad_v_narrow	2.70E-115	GGTMACAC	Motif_1.ohler	15.08440624	NA	TSS	0	3
1	GGYCACAC	all_tss	2.20E-198	GGYCACAC	Motif_1.ohler	14.92524926	NA	TSS	9.88131291682493e-324	3
1	GGYCACAC	broad_tss	1.20E-202	GGYCACAC	Motif_1.ohler	14.93886735	NA	TSS	9.88131291682493e-324	3
55	GTACTSS	internal_tss	6.6E-17	SSAGTAC	NA	11.77717969	NA	NA	NA	NA
27	GTATMTK	narrow_tss	2.9E-11	GTATMTK	NA	11.31979511	NA	NA	NA	NA
19	GTGGCMAC	broad_tss	1.1E-09	GTGGCMAC	NA	13.45576599	NA	TSS	1E-19	75
18	GWAAACA	all_tss	0.00000011	GWAAACA	NA	12.62824641	NA	TSS	6.4E-14	72
19	GYKGCCA	all_tss	4E-60	TGGCMRC	NA	11.93420615	NA	TSS	1.8E-34	52
16	HATCGATA	all_tss	9.40E-223	HATCGATA	DRE.ohler	14.24227177	BEAF-32	TSS	2.20E-200	183
16	HATCGATD	broad_tss	3.10E-249	HATCGATD	DRE.ohler	13.06002288	BEAF-32	Upstream	7.10E-183	183
23	HTGGCCA	narrow_tss	0.00084	HTGGCCA	NA	10.62837026	NA	TSS	3.6E-12	50
56	KCCTCCR	internal_tss	6.6E-16	YGGAGGM	NA	11.87979772	NA	NA	NA	NA
25	KGAAAAY	narrow_tss	0.0016	KGAAAAY	NA	11.63837871	NA	TSS	0.034	174
58	KTGGA	narrow_tss	0.000035	KTGGA	NA	9.00861146	NA	NA	NA	NA

4	MRCACACA	narrow_tss	0.00028	TGTGTGYK	NA	12.91253478	RNCMPT00178	NA	NA	NA
10	RACTRA	narrow_tss	9.60E-282	TYAGTY	INR.ohler	10.17863803	NA	TSS	0	3
43	RCAAWA	broad_v_narrow	2.5E-14	RCAAWA	NA	10.04843379	NA	NA	NA	NA
2	RDAAATA	all_tss	2.5E-24	RDAAATA	NA	11.4123739	NA	NA	NA	NA
43	RGCCAAR	narrow_v_broad	3.5E-11	RGCCAAR	NA	11.58570952	FBgn0037207	NA	NA	NA
54	RTATTTW	narrow_tss	0.00000023	RTATTTW	NA	11.52790356	NA	NA	NA	NA
32	RTCTGGCA	broad_v_narrow	0.03	RTCTGGCA	NA	12.76791716	NA	TSS	9.9E-18	47
44	RTGYA	broad_v_narrow	0.000000013	RTGYA	NA	8.047225816	NA	NA	NA	NA
11	STATAWA	narrow_v_broad	9.7E-101	STATAWA	TATA.ohler	11.85582409	NA	Upstream	1.8E-46	66
39	STGGA	internal_tss	2.80E-140	STGGA	NA	8.99432034	NA	NA	NA	NA
10	TACTGAA	narrow_tss	0.0000081	TTCAGTA	INR.ohler	11.95238118	NA	TSS	0	2
11	TATAWA	narrow_tss	0.033	TATAWA	TATA.ohler	10.36973332	NA	Upstream	0.00000013	65
11	TATAWAW	all_tss	9.10E-122	TATAWAW	TATA.ohler	12.0089178	NA	Upstream	2E-10	64
11	TATAWAW	narrow_tss	3.90E-190	TATAWAW	TATA.ohler	12.00112068	NA	Upstream	2E-10	64
16	TATCGAW	broad_v_narrow	1.50E-161	WTCGATA	DRE.ohler	12.99242019	BEAF-32	Upstream	7.10E-185	174
17	TCGWW	internal_tss	2.6E-62	TCGWW	NA	8.00658686	NA	Downstream	3E-92	144
25	TGGRAAW	all_tss	5.7E-14	WTTYCCA	NA	11.86337892	NA	NA	NA	NA
23	TGGVCR	internal_tss	7.00E-205	YGBCCA	NA	9.41662078	NA	NA	NA	NA
18	TTRTTTW	broad_tss	1.6E-92	WAAAYAA	NA	11.96368181	NA	NA	NA	NA
18	TTTRTTTW	all_tss	1.9E-84	WAAAYAAA	NA	13.8844452	RNCMPT00117	NA	NA	NA
17	WATCGR	broad_tss	3.9E-15	WATCGR	NA	9.982388836	NA	Upstream	5E-83	191
1	WGTGACCR	all_tss	1.8E-15	YGGTCACW	Motif_1.ohler	13.03665209	NA	TSS	1.70E-195	5

Table 3.4 – Discovered pA site Motifs

MotifName	readable_name	named_group	Positioned	E_value	DremeNum	TomTomMatch	centrimo_E_value	centrimo_adj_p_value	centrimo_bin_location
AHATAHAT	CPSF	Known	TRUE	1.40E-256	1	NA	0	0	-24
AATAMA	CPSF (Can)	Known	TRUE	1.80E-123	2	NA	0	0	-25

TKTGTDT	Cstf	Known	TRUE	1.30E-54	6	NA	0	0	-22.5
TGTACD	CFI/IIm	Known	FALSE	1.30E-26	9	NA	NA	NA	NA
AWATRTA	CPSF (2)	Known	TRUE	5.20E-21	12	NA	0	0	-24.5
CACWCAC	CACWCAC	Known	FALSE	1.90E-05	22	NA	NA	NA	NA
AAACSRA	AAACSRA	Unknown_proximal	TRUE	1.10E-24	10	unknown	0	0	-1.5
TKCABTT	TKCABTT	Unknown_proximal	TRUE	3.30E-17	14	unknown	0	0	7
GTTTYC	U2AF2	Unknown_proximal	TRUE	1.40E-10	16	unknown	0	0	11.5
RTATTT	SUP-26/SHEP	Unknown_proximal	TRUE	5.60E-07	19	unknown	0	0	-22.5
SAGCWG	SAGCWG	Unknown_distal	TRUE	2.00E-83	3	NA	0	0	-116
GGHGGM	GGHGGM	Unknown_distal	TRUE	5.50E-65	4	NA	0	0	-123.5
STGGV	STGGV	Unknown_distal	TRUE	1.30E-33	7	NA	0	0	-119
CRAGGA	CRAGGA	Unknown_distal	TRUE	6.70E-22	11	NA	5.80E-08	1.60E-09	-122
TCGA	TCGA	Unknown_Not_Positioned	FALSE	1.50E-57	5	unknown	NA	NA	NA
CARCAACW	YBX2/Unc75	Unknown_Not_Positioned	FALSE	2.40E-33	8	unknown	NA	NA	NA
SVAGAAG	Tra2	Unknown_Not_Positioned	FALSE	2.20E-20	13	unknown	NA	NA	NA
AAAAAAAA	A repeat	Unknown_Not_Positioned	TRUE	3.60E-19	15	NA	0	0	-15.5
GAYGAYGA	GAYGAYGA	Unknown_Not_Positioned	FALSE	9.50E-10	17	NA	NA	NA	NA
GAGMGAGA	GAGMGAGA	Unknown_Not_Positioned	FALSE	3.30E-09	18	NA	NA	NA	NA
CWACWAC	CWACWAC	Unknown_Not_Positioned	FALSE	8.50E-07	20	NA	NA	NA	NA
TGGCCA	TGGCCA	Unknown_Not_Positioned	FALSE	1.70E-06	21	NA	NA	NA	NA
AAGCGRA	AAGCGRA	Unknown_Not_Positioned	FALSE	2.30E-05	23	NA	NA	NA	NA
MCACCAYC	MCACCAYC	Unknown_Not_Positioned	FALSE	4.30E-05	24	NA	NA	NA	NA